Humanizing Digital Experiences: Three Essays on the Design of Digital Entities

A DISSERTATION

SUBMITTED TO THE FACULTY OF THE GRADUATE SCHOOL

OF THE UNIVERSITY OF MINNESOTA

BY

Scott D. Schanke

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

Dr. Gordon Burtch, Dr. Gautam Ray

May, 2021

Acknowledgements

First I thank my advisors Gautam Ray and Gordon Burtch.Our weekly meetings for these five years have always left me feeling invigorated with new ideas. Gautam, you opened my eyes to this wonderful world of research when I was a senior in undergrad here at Carlson, and without your encouragement I would never have known to start a PhD. Gord, I've learned so much from your approach to research, and your ability to sharpen research questions. As we three make a great team, I look forward to continuing our work together for years to come.

Additionally, a special thank you to my committee members Gedas and Brian, whose comments during this process have been invaluable. Gedas, you've taught me a lot about the technical aspects of machine learning and data mining over the years, and have also been a great sounding board and collaborator. Brian, your instruction in Computational Linguistics and Linguistic Pragmatics have served as the foundation for all the chapters in this dissertation.

I would also like to give a special acknowledgment to the wonderful faculty, staff, and PhD students in the IDSC department. The intellectual curiosity, fantastic seminars, and engaging discussions have provided fuel for me to keep pushing.

I also am grateful to the Minneapolis Business community, for being open to experimentation, and crazy ideas!

Next, I would like to thank my family for providing unconditional love, encouragement, comic relief, the occasional addition to my stipend, and a guest room in Milwaukee. Not to be forgotten my fur kids, Linus, Frankie, and Babygirl(you are missed) your need for food, walks, pets, and cuddles have given me ample distraction when needed.

Finally, I thank my wonderful fiancee Hudaa Mohammed. Your love, support, inquisitive mind, jokes, surprising interest in my research, and ear for concise writing have helped immensely throughout this process. I can't wait to continue our journey together in Milwaukee.

Abstract

Marketing and branding efforts have shifted from broadcast media, as in magazines or television, toward bi-directional/conversational media. Firm representatives are increasingly digital, and thus dynamic, autonomous and personalizable. Rooted in the shift in marketing practice, this dissertation seeks to identify and quantify effective approaches to the design and implementation of the entities that represent firms and brands in customer interactions, e.g., AI-enabled conversational agents, digital brand personalities. This thesis consists of three essays relating to the mediums in which digital entities exist (Social Media Pages, Messaging Applications, and Voice Based applications).

In my first essay on this topic, I evaluate how Politeness (Brown and Levinson 1987), a theory used to describe human request behavior, can be adapted to Social Media posts to further garner off platform sales conversions. This is important as it shows that the language used in Social Media posts are not uniformly perceived and can be tailored for customers depending on their relationship with the focal firm. The second essay moves from posts on social media, to messaging platforms. More specifically, in the context of customer service, I evaluate how the humanness of a conversational agent, (i.e. the number of social cues present),

influences customer service conversion outcomes, and customer price sensitivity. Our findings suggest that making an agent more humanlike can increase the rate of conversion for customers, however, customers also become more price sensitive in this particular "ultimatum game" like scenario. This shows that efforts to humanize conversational agents need to be carefully thought through and implemented to best support the context. In my final chapter, I explore the interactions between two key design choices for voice-based AI agents: i) disclosure of an agent's autonomous nature, and ii) aesthetic personalization (implemented via voice cloning). Through use of a Behavioral Economics game, we evaluate these features impact on trust. Overall, we find that people prefer a cloned version of an A.I. voice compared to a default male voice and no message control. Disclosure, on its own, does not significantly impact trust. When examining the interaction of message medium and agent disclosure, we find that dynamic voice cloning, in tandem with disclosure, achieves the highest user trust levels.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Years ago, communication between organizations and customers was unidirectional. These organizations would interact with their patrons via mediums like television, radio, and print media by sending messages into a void. Never could these expensive messages receive a real-time response. In the early 2000s, this unidirectional dynamic changed as social media rose to prominence. Organizations now could have an asynchronous conversation with their customers through their Facebook page or Twitter handles. As technology has advanced, now organizations can have extended text conversations through platforms like Facebook Messenger, WhatsApp, and WeChat. On the horizon, these interactions are even moving to voice-based exchanges through intelligent speakers and AI-enabled Interactive Voice Response Systems, (IVRs).

This burgeoning phenomenon, "conversational commerce," describes the ability for customers to receive customer support, make purchases, send payments, or receive product information through chat-based digital platforms.[1] Inundated with messages from their customers, organizations seek ways to automate their interactions through these platforms, as fully human-based interactions are infeasible. One way to do this is using AI-enabled conversational agents, also known as chatbots, that allow for an automated and instantaneous dyadic conversation. This year, Gartner even projects that 15% of all customer service will occur through chatbots, a 400% increase from 2017.[2]

While the first goal of automation focuses on mechanistic efficiency, organizations may lose sight of these conversational mediums' 'humanizing' aspects. This work seeks to better understand how to utilize elements present in human-to-human interaction and imbue them into digital entities representing organizations. More specifically, this dissertation evaluates how humanizing tactics in three digital customer touchpoints (Social Media, Messaging Platforms, and IVRs) influence economic outcomes of interest (e.g., purchase intention, price sensitivity, and trust), particularly relevant to conversational commerce.

---

[1] https://www.gartner.com/smarterwithgartner/the-art-of-conversational-commerce/
[2] https://www.gartner.com/smarterwithgartner/27297-2/

1.1 Essay Overview

My first essay on this topic focuses on the application of Politeness theory, developed in Linguistics, to explain how social relationships define individuals' approaches to making requests in various contexts. Politeness theory bears obvious relevance to marketing practice; consider that 'Buy Now', 'Click Here', or 'Follow Us' are directives that digital marketers frequently make of consumers via social media. Although these types of imperative statements (directives) (Searle 1975) are common on social media, prior work offers mixed evidence about their effectiveness in gaining customer compliance. Often, there is ample opportunity for miscommunication or unintended results to arise. For example, when messages are too assertive, customers may react negatively, and when they are too vague, customers may be confused.

A potential reason for the aforementioned mixed results is that customer relationship is not taken into account. In this work, we build on the conceptualization of Communal and Exchange relationships (Clark and Mils 1993, Aggarwal 2004) to understand how brand-customer relationships affect receptivity to different levels of Politeness. To do this, I partnered with a US-based women's clothing retailer to conduct a randomized field experiment on their Facebook page. In this experiment, I assigned consumers on Facebook to receive one randomly assigned social media advertisement. The ad-copy incorporated requests derived from alternative politeness strategies identified in Brown and Levinson's linguistic politeness theory (Brown and Levinson 1987). These requests ranged from implicit and indirect to explicit and blunt. Linking each consumer's advertisement exposure (and thus politeness level) to later compliance outcomes using the Facebook pixel, I arrive at several significant findings.

First, we show that explicit requests lead to significantly greater website conversion rates than implicit requests. Second, among explicit requests, consumers' response to increasing levels of politeness depends on their relationship to the brand, such that conversion probability is increasing in politeness among consumers who follow the brand's Facebook page (those in a Communal Relationship), and decreasing in politeness for consumers who do not follow the brand's Facebook page (those in an Exchange Relationship). The work offers experimental evidence that brands should phrase their requests to individuals on social media depending on their relationship with their audience, much like human-to-human interaction.

The second essay moves from posts on social media to messaging platforms and assesses the impacts of 'humanizing' AI-enabled autonomous customer service agents (chatbots). For this study, I developed and implemented a chatbot and ran a field experiment in collaboration with a dual-channel clothing retailer based in the United States. I worked with this retailer to automate a used clothing buy-back process, such that individuals would engage with the retailer's autonomous chatbot to describe (in chat) the used clothes they wish to sell, obtain a price offer, and (should they accept the offer) print a shipping label to finalize the transaction. We causally estimate the impact of chatbot anthropomorphism on transaction conversion by randomly exposing consumers to varied levels of anthropomorphism, operationalized by incorporating a random draw from a set of three anthropomorphic features: humor, communication delays and social presence.

This chapter provides evidence that anthropomorphism is generally beneficial for transaction outcomes, yet that it also leads to significant increases in customers' price sensitivity. We argue that the latter effect occurs because, as a chatbot becomes more human-like, consumers shift from a price-taking mindset into a fairness evaluation or negotiating mindset. We also provide descriptive evidence suggesting that the benefits of anthropomorphism for transaction conversion may derive, at least in part, from consumers' increased willingness to disclose the personal information necessary to complete the transaction.

These findings provide fascinating insights into the perils of humanizing A.I. performing human-facing customer service job roles. On the one hand, humanization can encourage higher levels of transactions; however, this may reduce the ability to adjust prices. The study is also one of the first field experiments capturing live user interactions with a chatbot.

The third chapter moves from studying text-based interaction to voice-based interaction. Voice-based media remain common customer service touchpoints (e.g., customer service call centers), and they are playing an increasingly significant role in digital sales (e.g., Amazon Alexa). As Interactive Voice Response agents (IVRs) become more commonplace in call centers and through smart speaker technology, it is crucial to understand how autonomous conversational agents' voice and speech characteristics may influence consumers' economic decision-making. In this work, I evaluate two crucial design characteristics of voice A.I. agents i) disclosure of an agent's autonomous nature and ii) voice personalization.

A.I. disclosure is becoming an important design feature as the current legal environment requires it. California recently passed the Bolstering Online Transparency Act (aka Bot Act or Blade-Runner Bill), which forces organizations to disclose the autonomous nature of their customer-facing A.I. Penalties for withholding this information at the onset can be up to $2,500 per interaction.[3] While transparency is often considered beneficial to a consumer's well-being, fieldwork evaluating A.I. disclosure in a sales scenario suggests that there may be drastic

[3] https://www.dwt.com/blogs/artificial-intelligence-law-advisor/2019/07/is-there-anybody-behind-that-bot

monetary consequences (Luo et al., 2019). As this disclosure policy may potentially become a federal regulatory requirement,future use of A.I. agents in human job roles may be contingent on designing interactions that mitigate these negative impacts.[4]

One feature that may help to reduce these concerns is the voice used by the A.I. agent. Notably, previous studies in Human-Computer Interaction and Psychology have investigated how similarity-based attraction manifests itself through Text to Speech technology, finding that individuals prefer agents with voice characteristics similar to their own (Dahlbäck et al. 2007, Nass and Lee 2001). That said, there may be ways that organizations can design the voice of these systems to create better user experiences. Interestingly new deep learning technology has enabled the ability to clone voices, in near real-time, with a short audio clip (Jia et al. 2018). This technology could be a viable means to personalize these interactions in the future.

In this work, we seek to answer the following research questions: ***To what extent does voice cloning induce trust in an autonomous (spoken) agent? Does disclosure of a spoken agent's autonomous nature causally impact user trust? How does disclosure affect the response to voice-based personalization?***

[4] https://www.congress.gov/bill/115th-congress/senate-bill/3127

To address these questions, we utilize a behavioral economics game initially used by Charness and Dufwenberg (2006), to evaluate how communication impacts trust. Disclosure likely influences trust, as disclosure may help to foster a perception of transparency to the user. Additionally, withholding disclosure may also have consequences if a user independently perceives the autonomous nature of the other party. In our study's version of the game, a human subject and an A.I. agent play a single round of the trust game. We randomly disclose the autonomous nature to the human player. After pairing with the agent, the human subject faces a choice of whether to trust the other party. Before the human subject makes their decision, the automated agent can randomly send a message. This message is communicated in either a dynamically cloned voice or a default male voice.

In our online game setting, we find that people prefer a cloned version of an A.I. voice compared to a default male voice and no message control. Notably, disclosure, on its own, does not significantly impact trust. When examining the interaction of message medium and agent disclosure, we find that dynamic voice cloning, in tandem with disclosure, achieves the highest user trust levels. These findings are interesting, as they suggest dynamic voice cloning could be a viable design feature for disclosed AI agents in the future.

Chapter 2

Experimental Evidence of the Effect of Politeness on Social Media Advertising Efficacy

2.1 Introduction

The last 10 years have witnessed the meteoric rise of social media. Where consumers once spent hours reading magazines or watching television, they now dedicate much of that time to consuming content on social media. On average, individuals spend an estimated 50 minutes each day interacting on social media platforms, a number that continues to grow[1] . Cognizant of this trend, marketers have diverted a large portion of their advertising efforts to targeting consumers on social media platforms in recent years. It is estimated that social media advertising spend now totals $36 Billion globally[2] .

Unlike print and television media, on social media, marketers are expected to take on an *anthropomorphic* persona and engage in conversation with consumer audiences (Aggarwal and Mcgill 2012). A common tactic employed by brands and marketers, to foster this perception of anthropomorphism, is to employ imperative verbs in the content and advertisements they post; that is, statements that explicitly request that the audience engage in a particular action, e.g. "Click Below" or "Buy Now" (Kwon and Sung 2011). In fact, it has been estimated that these sorts of imperative requests now appear in 44% of all Facebook posts generated by Interbrand's Top 100 Global Brands, and 29% of all Tweets generated by Business Week's Top 100 Firms (Chen et al. 2015, Kwon and Sung 2011). A significant body of work in print advertising (Moore et al. 2014), recommendation systems (Fitzsimons and Lehmann 2004), healthcare (Dillard and Shen 2005) and environmental sustainability (Kronrod et al. 2012a) has found that imperative

[1] http://www.adweek.com/digital/mediakix-time-spent-social-media-infographic/

[2] https://www.emarketer.com/Article/Social-Network-Ad-Spending-Hit-2368-Billion-Worldwide-2015/1012357

statements requesting action on the part of a target audience can lead to counter-productive outcomes, reducing an audience's intention to engage in the advocated behavior. However, the unique bidirectional nature of social media (Malthouse et al. 2013) implies a context in which imperative requests may in fact be permissible, if they are perceived as residing within the flow of natural conversation or social interaction.

The distinction between degree of assertiveness in implicit and explicit requests relates directly to the theory of politeness, in linguistics (Brown and Levinson 1987). Politeness theory is a sub-domain of linguistic pragmatics that considers the relationship between interpersonal factors and lexical decisions (word choice) on the part of interlocutors who are making requests of others (Holtgraves 2011). Importantly, this theory speaks not only to the coarse distinction between implicit and explicit, but also to nuanced gradations in the level of politeness that individuals employ to elicit a desirable response from a conversation partner. This notion of politeness is clearly relevant to the domain of advertising, once advertisements are viewed requests are made of consumers. Notably, to our knowledge, the notion of politeness has received limited attention in past work. Accordingly, we leverage this theory, in an effort to gain a more nuanced understanding of how shopping requests should be phrased, not only in terms of explicitness, but in terms of politeness more generally.

Thus, in this work, we address the following key research questions: ***First, does explicitness increase or decrease the efficacy of advertisers' online shopping requests on social media? Second, how and to what degree does the efficacy of advertisers' shopping requests depend on the politeness of a request? Third, to what degree do these relationships depend on consumers' relationship with the advertising brand?***

We evaluate these questions by partnering with a US-based dual-channel retailer of women's clothing that maintains a strong social media presence. We leverage that presence to conduct a randomized field experiment on Facebook, randomly assigning customers to receive alternative social media advertisements that vary in their level of politeness. In particular, we linguistically manipulate a typical imperative advertisement that asks a consumer to shop at the retailer's online store, utilizing Facebook's native A/B split-testing features to evaluate the effectiveness of alternative advertisements. We then link a consumer's exposure to a particular advertisement condition on Facebook with any online shopping events that a consumer makes on the retailer's website over the following seven days, utilizing the Facebook Pixel (Gordon et al. 2017).

We document a number of useful findings. First, we observe that explicit shopping requests by this retailer drive 13x more website conversions than implicit requests. This result is interesting because it is seemingly contradictory to previous linguistic and advertising studies, which collectively suggest that these types of assertive statements are not effective (Dillard and Shen 2005, Fitzsimons and Lehmann 2004). Delving deeper, however, we provide evidence that the efficacy of explicit requests, relative to implicit requests, actually depends on multiple factors, including the level of politeness in the explicit request, and the targeted consumer's pre-existing relationship with the brand on social media. Specifically, we find that less polite language directing the consumer to take immediate action works well for viewers that are not connected to brand's Facebook page, while more polite language is effective for users connected with the brand's Facebook page. We conclude that the reason for these diametrically opposing results is that consumers, when engaged in a relationship with a brand, have different norms and expectations depending on the type of relationship (Aggarwal 2004). We theorize that when relationships are motivated solely by economic transactions i.e., *Exchange Relationship* (Aggarwal 2004), politeness in a request is superfluous as the norms of the relationship between brand and customer are impersonal and focus solely on exchange of goods and services. In contrast, when relationships place emphasis on social interactions and go beyond mutual self-interest i.e., *Communal Relationship* (Aggarwal 2004), emphasizing more polite language in a request shows to the customer that the brand values the relationship, and transcends self-interest.

Our work contributes to marketing literature (Fitzsimons and Lehmann 2004, Zemack-Rugar et al. 2017), by further investigating how word choice influences compliance. To the best of our knowledge, our work is the first to evaluate the efficacy of brand-generated imperative requests in a social media setting. We also contribute to marketing literature (Moore et al. 2014) as this study further explores the use of assertive messaging, and evaluates its effectiveness on customers with varying degrees of relationship distance.

We contribute to digital marketing literature (Bapna et al. 2017, Lee et al. 2017, Schanke 2017) by investigating novel aspects firms should consider in their content engineering efforts on social media. Where previous studies have looked at on-platform engagement, comment valence (Schanke 2017), general engagement (Lee et al. 2017) and community growth (Bapna et al. 2017), this work considers how firms' messaging tactics directly impact an off platform outcome, website conversions. Whereas previous studies have investigated features of social media posts that moderate their efficacy at engaging a firm's audience in aggregate (Lee et al. 2017, Schanke 2017), this study highlights how language used in Social Media posts can be perceived differently when viewed by people in different relationships with the firm.

2.2 Literature Review

2.2.1 Social Media Marketing

Miller and Tucker (2012) offer an early example of work studying the value of firm engagement on social media. These authors report evidence that a firm's decision to take active ownership of its online social media presence yields greater consumer engagement. Goh et al. (2012) examine the value organizations derive from different types of social media content, comparing that produced by users with that produced by the firm. They find that both types of content have an impact on firm sales, with user-generated content being substantially more influential.

Lee et al. (2017) drill down to investigate the features of specific pieces of content that brand marketers post to their Facebook pages, to understand how different features associate with social engagement and click-through. These authors find that, when employed in isolation, brand personality features, such as humor or emotion, yield greater social engagement than informational features, such as price, whereas informational features yield greater click-through. However, when the two are combined, this yields the best overall engagement. These findings speak to the importance of content features, and the broad opportunity of content engineering to facilitate consumer engagement and response.

Whereas Lee et al. (2017) considered a wide variety of companies and brands in their study, making general claims about about the average returns to content features, Schanke (2017) examined heterogeneity in the returns to content features across different product pages on Facebook, characterizing products in terms of two aspects: hedonic (vs. utilitarian) and high (vs. low) involvement (Schanke 2017). These authors find evidence that these characteristics moderate engagement with language features (Formal vs Informal) and Business vs Non-Business content. Moreover, they find that incongruous posting (e.g. Hedonic firms posting business content, when they typically post non-business content) is met with higher levels of engagement.

The present study contributes to this stream of literature by considering how a specific form of brand-posted content on social media, i.e., imperative requests, influences website conversion, and how that relationship is jointly moderated by the degree of politeness employed in the firm's requests (advertisements) and the audience's relationship with the brand.

2.3 Theory and Hypotheses

2.3.1 Explicit vs. Implicit Requests

Imperative requests have been a mainstay of marketing for decades. Such requests commonly appear in direct marketing mailers, print advertisements, and online advertisements (Applegate 2016). Examples of this include prompts like: "I urge you to act at once", "Order now, while there's still time" or "Send in your application today!" (Bayan 2006). These statements are clear, and also highlight a sense of urgency to elicit a rapid response (Danaher et al. 2015). On the surface, it seems likely that the clarity and directness of these requests would make them particularly effective at eliciting the desired consumer response; however, these types of solicitations have met with mixed success in practice (Miller et al. 2007, Moore et al. 2014, Dillard and Shen 2005, Kronrod et al. 2012a).

On the one hand, some prior work has observed that messages utilizing clear and concrete language are better able to attract the attention of targets than those employing abstract language (Miller et al. 2007). Concrete messages give the message receiver clear direction and details, thereby reducing the potential for confusion or mis-communication (Miller et al. 2007, Holtgraves and Yang 1990). Additionally, these prompts have been found to work particularly well when the recipient of the message is in a good mood (Kronrod et al. 2012a) or cares about the 'cause' (Kronrod et al. 2012b).

However, directly demanding a response has also been shown to elicit *reactance* on the part of consumers (Miller et al. 2007, Brehm 1966, Zemack-Rugar et al. 2017), a psychological phenomenon in which individuals feel that their freedom or independence is under threat. This perception of being controlled or manipulated can elicit negative responses, such as anger, frustration and demotivation, which in turn lead to non-compliance (Dillard and Shen 2005, Zemack-Rugar et al. 2017, Deci and Ryan 1985). Because of this, advertisers often opt instead for more passive, indirect language, such as “now is a good time to buy”, and “always fresh” (Zemack-Rugar et al. 2017), wherein the request to purchase is implicit in nature. Indirect requests have the potential to be more effective at inducing a desirable response, primarily because they place less pressure on a consumer to comply.

Ultimately, the dominant strategy in the specific context of social media advertising remains an open empirical question, given the conflicting evidence across alternative media that has been documented in the prior literature. Accordingly, we begin with the following pair of competing hypotheses about the relative efficacy of explicit versus implicit requests:

For the General Facebook Audience, making an explicit shopping request will lead to greater website conversion compared to an implicit shopping request.

For the General Facebook Audience, making an implicit shopping request will lead to greater website conversion as compared to an explicit shopping request.

2.3.2 Politeness

We should begin by noting that the theory of politeness is both descriptive and normative in nature (Brown and Levinson 1987). It was formulated based on extensive observation and study of face-to-face interpersonal communication and conversation in offline settings. Because it is normative, the theory represents something of an ideal for dyadic, interpersonal conversation. Here, we apply politeness theory in a prescriptive fashion, to theorize how marketers should make requests of their audience in online environments. Importantly, deviation from the conversational ideal is particularly likely to occur in online settings, because dyadic interactions are constrained by a lack of media richness (Daft and Lengel 1986), synchronicity and social cues (Kiesler et al. 1984). Because of these constraints, normative interpersonal communication practices are more likely to be misapplied and misinterpreted on social media.

Politeness theory describes linguistic strategies that individuals employ when attempting to elicit cooperation from others. The theory speaks to various politeness strategies, which are employed by speakers under different circumstances, depending on the nature of the interpersonal relationship between the speaker and the recipient, and the nature of the request being made. Brown and Levinson (1987) root their theory of politeness in the sociological concept of 'face' (roughly related to public self-image, or self-esteem), a concept first developed by Erving Goffman Goffman (1967). Face, which may be lost, gained or maintained through interactions with others (Jaworski and Coupland 2014), has two types: negative and positive. Negative face refers to the personal need for independence and the freedom to carry out actions without imposition (Brown and Levinson 1987, Holtgraves 2011). Positive face, on the other hand, refers to the desire for public, social approval (Brown and Levinson 1987). In politeness theory, requests and impositions specifically threaten a hearer's negative face. Notably, there are clear parallels between the notion of negative face and the concept of reactance (Dillard 2007).

Politeness theory suggests that, in any social interaction, the two types of face are inherently threatened (Holtgraves 2011). The degree to which one's statements threaten others is determined by three factors: the power distance between the speaker and recipient, the social distance between them, and the degree to which a request imposes on the recipient (Brown and Levinson 1987, Holtgraves 2011, Jaworski and Coupland 2014). As the power distance, social distance or degree of imposition increase, the potential for face threat grows in turn (Holtgraves and Yang 1990, Jaworski and Coupland 2014). Politeness theory argues that speakers always employ one of four request strategies, depending on the interpersonal factors and the nature of the request: *Bald on the record, Positive*

Politeness, Negative Politeness, and Off the Record (each strategy is described in detail in the section below).

Politeness theory suggests that speakers typically make requests in a manner that depends on the interpersonal variables noted earlier, i.e., social distance, power distance and degree of imposition (Brown and Levinson 1987, Holtgraves and Yang 1990). If the speaker and the listener are socially proximate (e.g. close friends) then the style of interaction is typically more direct and lacks politeness (Holtgraves 2011, Brown and Levinson 1987, Blum-Kulka 1987). Conversely, if a listener holds a greater degree of relative power in the relationship, or if the speaker's request would amount to a large imposition on the listener, greater politeness is typically employed (Brown and Levinson 1987).

Among the three factors mentioned, here we focus on social distance, which bears perhaps the most obvious relevance. We operationalize social distance between a consumer and a brand based on a consumer's pre-existing engagement with the brand's online community. We argue this operationalization in the following section, based on the notion of communal and exchange relationships.

2.3.3 Communal vs. Exchange Relationships

Although traditional politeness theory focuses on language use in utterances between individuals, the present study seeks to understand language use in utterances from a brand to an individual. Human relationships with brands are somewhat different from interpersonal relationships, but there are also important similarities (Aggarwal and Mcgill 2012, Aggarwal 2004, Fournier 1998). One shared conceptualization that applies to both interpersonal (Clark and Mils 1993) and consumer-brand interactions (Aggarwal 2004) is the distinction between a communal or exchange relationships.

Communal relationships are those based primarily on social factors (Clark and Mils 1993). Individuals in communal relationships seek to transcend self-interest, and provide benefits to others without any expectation of repayment (Clark and Mils 1993). Prototypical examples of this include relationships involving best friends, family members, or romantic partners (Aggarwal 2004). Analogous examples of such relationships also manifest in context of consumers and brands, when consumers engage socially in brand communities (Aggarwal 2004).

Exchange relationships are relatively asocial, being relationships that are largely transactional in nature (Clark and Mils 1993). A relationship is characterized as exchange-based when the parties involved are socially distant from one another, interacting solely for the sake of economic exchange (Clark and Mils 1993). The norms underpinning exchange-based relationships are quid-pro-quo (Aggarwal 2004). Individuals in these relationships expect to receive in-kind returns for any investments they make in the relationship. Relationship participants also typically maintain a formal or mental balance of accounts with the other party, as well, which they aim to hold even (Aggarwal 2004). Like communal relationships, exchange-based relationships are also common in consumer-brand settings (indeed, exchange-based relationships are most common in this setting). Concretely, such relationships would arise when consumers purchase a brand's products or services, but make no attempt to engage socially with the brand's social media presence or online community.

2.3.4 Expected Interaction Between Relationship and Politeness

Brown and Levinson (1987)'s politeness model has not been directly tested in a marketing context, nor has it been considered in tandem with the notion of or Exchange relationships. However, Yoon et al. (2016) suggest that utilizing a more polite strategy places emphasis on the social utility between interlocutors.

In our setting, we theorize that utilizing more polite strategies will potentially place emphasis on the social utility of the relationship between the page and customer. It is with this investment shown through more polite language that the firm signifies that it values the customer, and does not simply seek to profit off of the relationship. This is in a similar vein to the findings of (Zemack-Rugar et al. 2017), which find that less assertive, more polite language is received better by those in a close relationship with a brand. This, we believe, is more congruent with the norms of a Communal Relationship, thus we propose the following hypothesis:

For socially connected consumers (i.e., those in a Communal Relationship) the most polite explicit request (Negative Politeness) will lead to higher compliance as opposed to the less polite (Bald on the Record, Positive Politeness) and control (Off the Record) requests.

Contrasting the above, it is not necessarily clear how politeness will impact website conversion when advertisements are directed toward consumers in an exchange-based relationship with the brand. When the social distance between a speaker and recipient is greater, politeness theory suggests that individuals would employ more polite language (Brown and Levinson 1987). Although Zemack-Rugar et al. (2017) found that customers in committed brand relationships react negatively towards assertive advertising, they also found that participants in un-committed brand relationships preferred advertisements that were less assertive.

Despite these findings, Aggarwal (2004), contends that exchange based relationships are characterized by both parties being less concerned with each other's emotional states. As the politeness level is the lowest for Bald on the Record requests, it would follow that this messaging strategy would be most aligned with exchange relationship norms. This leads us to the following hypothesis:

For website visitors not connected to the page (Exchange Relationship), the least polite language (Bald on the Record) will lead to higher compliance as opposed to more polite requests (Positive Politeness, Negative Politeness and Off the Record)

2.4 Empirical Strategy & Analysis

2.4.1 Experiment Design

To evaluate these hypotheses we partnered with a US-based dual channel retailer of women's clothing to execute a between subject randomized field experiment, inserting our manipulations into the firm's regular social media marketing activity. We worked with the retailer to craft four distinct versions of the same advertisement, promoting its online store. The four versions of the advertisement held subject matter constant, but delivered it using alternative phrasings, wherein the politeness inherent in the shopping request was manipulated in line with the strategies documented in the literature on politeness theory (Bald on the Record, Positive Politeness, Negative Politeness, and Off the Record / Implicit). Similar to Gordon et al. (2017), we utilized the Facebook split-test advertising feature to randomly assign consumers into one of four experimental conditions, where each condition was exposed to one of the aforementioned versions of the advertisement. We then monitored subsequent website activity traced back to customers in each condition. We were able to trace customer website conversion to Facebook advertisement exposure utilizing the Facebook conversion pixel, which was installed on the retailer's online store website[3] .

Subjects in the experiment were drawn from two populations of Facebook users. The first population was the entire set of the firm's Facebook page followers; that is, consumers who had liked the retailer's Facebook brand page at some point prior to the experiment, and thus who held a prior social relationship with the brand (consumers in a Communal Relationship with the brand). This population included approximately 43,000 individuals. The second population was comprised of other Facebook users who had visited the retailer's online store within the prior 180 days, yet who also were *not* followers of the Facebook brand page (consumers in an Exchange Relationship with the brand). This population included approximately 20,200 individuals. To identify the latter population, we

[3] https://www.facebook.com/business/help/952192354843755

again relied on the Facebook Pixel, installed on the retail website, which enabled the firm to link website visitors to Facebook user accounts.

To ensure our experiment population was representative of the broader population of Facebook users interacting with the brand's advertisements and content on Facebook, we enforced a consistent 1:2 ratio of socially-connected:non-connected users across all four experimental conditions. This experimental design enabled us to causally identify the relative efficacy of request-based posts under exogenously varied politeness levels, and across Communal versus Exchange relationship types, while holding power distance and degree of imposition constant.

Many firms utilize the Facebook Feed algorithm (a.k.a. Edge Rank) to intelligently target users that are predicted to be quite likely to engage with a brand's post. However, these content-targeting features are problematic when it comes to conducting an experiment on Facebook, because different content may be promoted differently, yielding different levels of engagement and conversion that cannot be attributed solely to differences in the content itself (differential outcomes may also derive from imbalance in the distributional characteristics of the audiences exposed) (Eckles et al. 2018). Fortunately, Facebook's split-test tool enables advertisers to optimize solely on impression volumes, disabling engagement-optimizing targeting algorithms. Importantly, it is a relatively straightforward matter to account for the number of advertisement impressions in subsequent analyses.

2.4.2 Independent Variables

Once randomized into an experimental condition, each subject was then exposed to only one of the four advertisement versions over the subsequent 7-day period. That is, if a subject was assigned to a particular advertisement version, whenever the subject encountered the retailer's advertisement over the course of the subsequent week, they would always see the same version.

As noted earlier, each of the advertisement versions incorporated features that aligned with the request strategies proposed by Brown and Levinson (1987). The request strategies, along with associated linguistic features, are presented in Table 3.2. The exact text used in each treatment condition advertisement is then presented in Table 2.2.

Table 2.1: Message Features Used for Each Request Strategy

Request Strategy	**Description**	**Message Feature**
Bald on the Record	*Maximally Efficient Communication*	Urgency
Positive	*Claim common ground and Shared understanding*	Slang
Negative	*Be indirect / Non-imposing with the request*	Distance / Nominalization
Off the Record	*Imply a request*	Hinting

Bald on the Record – This politeness strategy does not use any form of linguistic politeness that would limit the face-threatening act. Rather, with this strategy, the speaker states the face threatening act in the clearest of terms (Holtgraves

Table 2.2: Request Strategy Message Content with Politeness Level

Request Strategy	**Politeness Level**	**Post Message**
Bald on the Record	1 (least polite)	New Styles. Shop now!
Positive	2	New Styles, shop 'til ya drop!
Negative	3	Consider shopping new styles
Off the Record	4 (most polite)	New Styles

2011), i.e., in a blunt fashion. To operationalize this strategy in our treatment message, we utilize the adverb *now*, with an exclamation mark.

Positive Politeness – Utilizing a positive politeness strategy, a speaker seeks to build trust, friendship and a rapport with the listener, thereby mitigating face threats and inducing the listener to perceive himself or herself as a member of the speaker's in-group. Positive politeness can best be described as an attempt to seek solidarity or common ground with the listener (Holtgraves and Yang 1990). This strategy implements the weakest form of politeness on the Brown and Levinson (1987) scale. Consistent with the politeness literature, we operationalize this strategy in our treatment message by employing slang, using both contractions and abbreviations, *'til, ya*, and more generally a colloquialism, *Shop 'til ya drop.*

Negative Politeness – Utilizing a negative politeness strategy, a speaker tries to avoid directly imposing on the listener, while still making clear the intent. With this tactic, the speaker seeks to respect the listener's desire for independence and freedom, and to be unencumbered by requests (Holtgraves and Yang 1990). This strategy is viewed as relatively more polite than the *positive politeness*

Figure 2.1: Bald on the Record Advertisement

strategy (Brown and Levinson 1987). Consistent with the politeness literature, we operationalize this strategy in our treatment message by incorporating the verb *Consider*, which shifts the listener's focus to "thinking" about shopping instead of actually shopping. Additionally, we incorporate a nominalization, which modifies the verb *Shop* into its gerund form, *Shopping.* This passive construction also places distance between the firm and the request being made.

Off the Record – The Off-the-record politeness strategy is the most polite of the four (Brown and Levinson 1987), and the least direct. This form of politeness does not involve an explicit request; rather, it relies on indirect implication (Jaworski and Coupland 2014). For example, a speaker in a room with an open window might say to a listener: "it's cold in here." From this statement, the speaker might hope that the listener would interpret the implied request that he or she close the window. To operationalize this strategy in our treatment message, we omit any reference to shopping, and simply make reference to *New Styles* being available. This treatment, which lacks any formal request, and which is most polite, serves as our reference (control) condition in our subsequent analyses.

Manipulation Checks

Upon constructing the treatment messages for each experimental condition, we performed a series of manipulation checks via Amazon Mechanical Turk, recruiting 28 workers to evaluate the language of the advertisements, in a manner similar to Holtgraves and Yang (1990). This was done to ensure that our chosen wordings resulted in the desired politeness perceptions among readers. Turkers evaluated the politeness of each advertisement version using a 7-point Likert-scale. In the survey, the respondents were asked:"Rank how bossy you deem each company's Facebook post on a 7 point scale, with 1 being extremely bossy and 7 being definitely, not at all bossy." Each Mechanical Turk worker was paid a total of $4.50 to evaluate the four manipulations. Evaluations were obtained from all 28 respondents for three of the manipulations (Bald on the Record, Positive Politeness, and Negative Politeness), and from 26 respondents in the case of Off-the-Record (because 2 workers did not provide the last evaluation). Descriptive statistics from the Mechanical Turk evaluations can be seen in Table 2.4. The results of the statistical comparisons between each pair of conditions using a Wilcoxon Sign Rank test are presented in Table 2.3. Note that the *p*-values were adjusted based on the Benjamini-Hochberg technique at the .05 level, to account for multiple comparisons. Despite this relatively conservative adjustment, we observe statistically significant differences in perceived politeness across all comparisons, in the expected directions. The only exception here is that there is no statistically significant difference between the messages used in our Negative Politeness and Off-the-Record treatments, suggesting that Mechanical Turk workers perceived the two to be equivalent. However, this result is actually consistent with other prior work Holtgraves and Yang (1990).

Table 2.3: Results Wilcoxon-Sign Rank test for Manipulation's Perceptions of Politeness

Condition Comparison	**Prob >z**
Bald vs Positive	0.015
Bald vs Negative	0.001
Bald vs Off	0.001
Positive vs Negative	0.008
Positive vs Off	0.003
Negative vs Off	0.682

Table 2.4: Manipulation Check Descriptive Statistics

Message	**Obs**	**Mean**	**Std. Dev.**	**Min**	**Max**
Bald on the Record	28	3.714286	1.843048	1	7
Positive	28	4.392857	1.770944	1	7
Negative	28	5.642857	1.282771	3	7
Off the Record	26	5.8	1.433661	2	7

2.4.3 Dependent Variable

To relate a subject's website conversion on the retailer's website to advertisement exposure on Facebook, we utilize the Facebook conversion pixel. A website conversion event was logged for this study if a subject viewed a product page, searched the website or added an item to their cart. Facebook's attribution of a website conversion occurs only if the customer makes an action within 24 hours of viewing the advertisement. If the a customer clicks a link embedded in the advertisement, then the attribution of activity for that customer's impression is open for 1 week. This is the default conversion attribution system used by Facebook for ads. Additionally, a singular impression can only be linked to one conversion event attributed to the advertisement.

It is important to note that the Facebook pixel is capable of linking this online activity to an advertisement impression even when the subject switches devices (e.g,. if the consumer sees the advertisement on their phone and then makes converts using a laptop)[4] .

2.4.4 Analysis Approach

We first examine group-level differences in conversion rates across the four experimental groups, employing two-tailed t-tests and Chi^2 tests. We thus perform six pairwise comparisons, accounting for the False Discovery Rate using a Benjamini-Hochberg adjustment at $\alpha = .05$. Subsequently, we separate our data into the two

[4] https://www.facebook.com/business/news/cross-device-measurement

sub-populations of interest (consumers holding Communal versus Exchange relationships with the brand), repeating our analyses, in a similar fashion. To assess robustness, we also subsequently perform an ANOVA test, as well as frequency-weighted Logistic Regression and a set of permutation tests.

2.5 Results

The group-level results are reported in Table 2.5. Reach is the number of subjects per condition, and impressions is the number of times the ads were viewed. On average, individuals saw ads about 8 times over a week long period. Also, we observe that the highest average conversion rate is associated with the *Bald-on-the-Record* treatment, with a website conversion rate of .005416, or 0.54%. The lowest response is associated with the *Off-the-Record* treatment, the most polite condition (Brown and Levinson 1987), and our control group. We also observe a standard deviation more than 10 times the mean in each condition, suggesting a great deal of variation in subjects' response, which is quite common for digital advertising campaign conversion and click through data (Goldfarb 2015).

Table 2.5: Conversion Rate by Condition

Condition	**Reach**	**Impressions**	**Mean**	**Std. Dev**	**Min**	**Max**
Bald on the Record	641	4801	0.005416	0.073192	0	1
Positive	617	4633	0.002158	0.046359	0	1
Negative	648	4909	0.004482	0.066644	0	1
Off the Record	619	4928	0.000406	0.020137	0	1

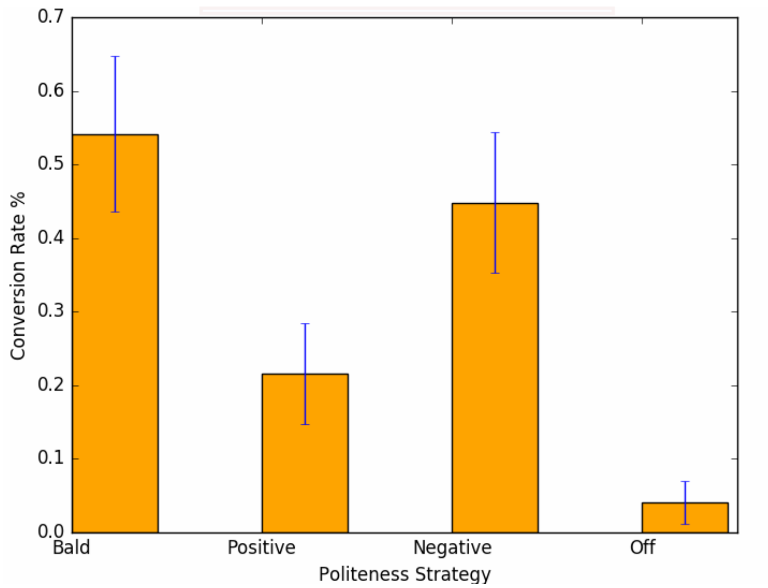

Figure 2.2: Results Overall Campaign

Table 2.6: Benjamini-Hochberg Adjusted Pairwise Comparisons

Test	**Condition Comparison**	***t*-stat**	***p*-value**	$\boldsymbol{Chi^2(1)}$	***p*-value**
1	Bald vs Off	4.6287	0.000	21.1443	0.000
2	Positive vs Off	2.4173	0.019	5.8364	0.024
3	Negative vs Off	4.102	0.000	16.7033	0.000
4	Bald vs Positive	2.5725	0.015	6.5302	0.022
5	Bald vs Negative	0.6578	0.511	0.4262	0.514
6	Positive vs Negative	-1.9666	0.059	3.8232	0.061

Figure 2.2 depicts the conversion per impression for each condition as a percentage, with the standard errors represented as error bars. As noted above, the politeness strategies that performed best were the *Bald on the Record* strategy (least polite), and the negative politeness strategy (theoretically, the second most

polite manipulation, though our manipulation checks suggest it is indistinguishable from the Off-the-Record manipulation, in a statistical sense). That said, the responses to the *Bald on the Record* and *Negative Politeness* strategies are not statistically distinguishable from one another. Pairwise comparisons, using t-tests and Chi-square tests, are presented in Table 2.6.

Considering our first pair of competing hypotheses (related to the relative efficacy of explicit and implicit shopping requests), we found evidence that explicit requests were significantly more effective than implicit requests; all three explicit request manipulations outperformed the implicit, Off-the-record manipulation. All differences were significant at the $p \leq .05$ level. We therefore found evidence in support of Hypothesis 1a, and counter to Hypothesis 1b.

Next, we considered the moderating effect of consumer-brand relationship (Hypotheses 2 and 3). We evaluated each hypothesis using a sub-sample analysis, separately considering data on subjects who had liked the brand Facebook page (Communal Relationship), and those who had visited the retailer's online store but had not liked the brand Facebook page (Exchange Relationship). Table 2.7 reports descriptive statistics around conversion rates across our experimental conditions, by relationship type. These results indicate that the *Negative Politeness* strategy was most effective among consumers in a Communal Relationship with the brand, whereas the *Bald-on-the-Record* strategy was most effective among consumers in an Exchange Relationship with the brand.

Table 2.7: Conversion Rates by Condition and Relationship

Condition	**Relationship**	**Reach**	**Impressions**	**Mean**	**Std Dev**	**Min**	**Max**
Bald on the Record	Communal	434	3196	.001877	.043247	0	1
Positive Politeness	Communal	418	3025	.000992	.031461	0	1
Negative Politeness	Communal	440	3144	.006043	.077269	0	1
Off the Record	Communal	407	3164	0.000	0.000	0	1
Bald on the Record	Exchange	207	1605	.012461	.110238	0	1
Positive Politeness	Exchange	199	1608	.004353	.065692	0	1
Negative Politeness	Exchange	208	1765	.0017	.041158	0	1
Off the Record	Exchange	212	1764	.001134	.033634	0	1

Figure 2.3 & 2.4 presents visual depictions of the conversion rates that were observed among consumers in Communal and Exchange relationships with the brand, by experimental manipulation, with standard errors. Pairwise comparisons based on t-tests and Chi^2 tests are reported in Table 2.8. In the Communal Relationship sub-sample, we found that the *Negative Politeness* strategy yielded conversion rates significantly higher than the *Bald-on-the-Record* strategy at $p = 0.016$, the *Positive Politeness* ($p = 0.003$) strategy and the *Off-the-Record* strategy ($p < 0.001$). These findings provide support for Hypothesis 2. In contrast, in the the Exchange Relationship sub-population, we found that the *Bald-on-the-Record* strategy yielded conversion rates significantly higher than *Positive Politeness* strategy ($p = 0.026$), the *Negative Politeness* strategy ($p < 0.001$), and the *Off-the-Record* strategy ($p < .001$). These results provide support for Hypothesis 3.

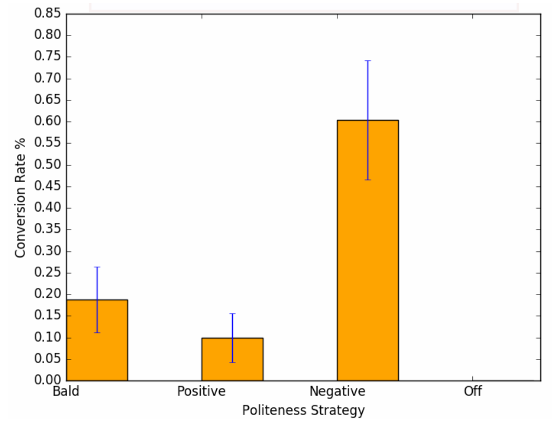

Figure 2.3: Conversion Rate, Communal Relationship

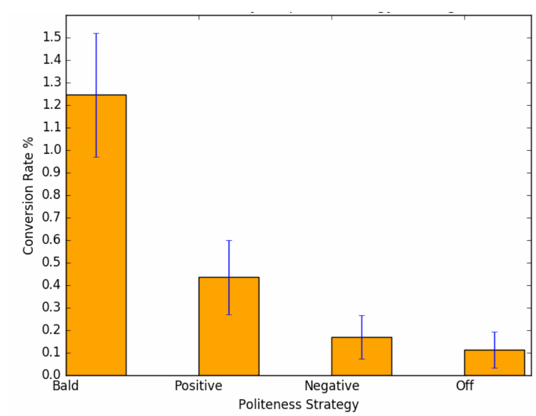

Figure 2.4: Conversion Rate, Exchange Relationship

Table 2.8: Benjamini-Hochberg Adjusted Pairwise Comparisons by Relationship

Condition Comparison	**Relationship**	***t*-stat**	***p*-value**	***Chi*^{2}(1)**	***p*-value**
Bald on the Record vs Positive Politeness	Communal	0.9187	0.358	0.8413	0.359
Bald on the Record vs Negative Politeness	Communal	-2.6545	0.014	6.9478	0.016
Bald on the Record vs Off the Record	Communal	2.4413	0.021	5.9344	0.022
Positive Politeness vs Negative Politeness	Communal	-3.3389	0.002	10.9933	0.003
Positive Politeness vs Off the Record	Communal	1.7736	0.089	3.1363	0.092
Negative Politeness vs Off the Record	Communal	4.4015	0.000	19.0633	0.000
Bald on the Record vs Positive Politeness	Exchange	2.533	0.023	6.2313	0.026
Bald on the Record vs Negative Politeness	Exchange	3.8188	0.000	14.16	0.000
Bald on the Record vs Off the Record	Exchange	4.1104	0.000	16.3968	0.000
Positive Politeness vs Negative Politeness	Exchange	1.4184	0.187	1.9922	0.19
Positive Politeness vs Off the Record	Exchange	1.8137	0.105	3.2575	0.106
Negative Politeness vs Off the Record	Exchange	0.4473	0.655	0.1992	0.655

2.5.1 Robustness Checks

We next evaluated the robustness of our results by implementing our analyses in a number of alternative ways, including ANOVA, Logistic Regression and permutation tests. The results of these additional analyses follow.

ANOVA

We began by performing a one-sided ANOVA test on the aggregated means, as well as the aggregated means by sub-population (Communal vs. Exchange consumers). Our results considering the whole population yielded consistent results with our main analysis ($F = 8.10$, $p \leq .001$). This result indicates, specifically, that the aggregate means across treatment groups are significantly different from one another. We obtained similar results in the Communal ($F = 10.01$, $p \leq .001$) and Exchange ($F = 9.66$, $p \leq .001$) sub-populations.

Logistic Regression

In Table 4.4 we report the results of a logistic regression of conversion on our treatment dummies. This regression is performed on the aggregate data, where the group-level observations are frequency weighted based on the Facebook reported impression counts for each campaign[5] . Because conversion is extremely rare and

[5] https://www.stata.com/support/faqs/statistics/logistic-regression-with-grouped-data/

standard logistic regression models are therefore likely to yield a biased estimate (Ye et al. 2014), we employ the rare-events (Firth) logit estimator (Firth 1993).

Overall, the results indicate that the *Bald-on-the-Record* treatment is significantly more effective than the *Off-the-Record* treatment (the omitted group) at $p \leq .01$. Also the Negative Politeness treatment is more effective than the *Off-the-Record* treatment, this is significant at the $p \leq .01$ level. Interestingly, the Positive Politeness also appears to perform significantly better than the *Off-the-Record* treatment, this is significant at the $p \leq .05$ level.

Table 2.9: Logistic Regression Overall Campaign Results

DV = Conversion	**Coefficient**
Bald on The Record	2.3920*** (0.6619)
Positive Politeness	1.4986** (0.7040)
Negative Politeness	2.2052*** (0.6670)
Observations	19,271

*** $p \leq .01$, ** $p \leq 0.05$

We further investigate the moderating effect of relationship type on the effectiveness of the politeness treatment. While in typical OLS regression, one would simply look at interaction terms, because we are using a nonlinear model to estimate conversion, interaction terms are considered to be biased (Ai and Norton 2003). Thus we run conditional logistic regressions for each sub-population, those which are connected to the page and those that are not.

First we consider those which are in a *Communal* relationship, Table 2.10. As can be seen, the Negative Politeness treatment performed significantly better than the *Off-the-Record* treatment control group. Interestingly, the *Bald-on-the-Record* treatment performed better than the control group, but the significance is far less pronounced than in the Overall Campaign results with a ($p \leq .10$). The significance associated with the Positive Politeness condition, also vanishes in this subgroup.

Table 2.10: Logistic Regression Results Communal Relationship

DV = Conversion	**Coefficient**
Bald on The Record	2.564* (1.468)
Positive Politeness	1.992 (1.512)
Negative Politeness	3.676*** (1.432)
Observations	12,529

*** $p \leq .01$, ** $p \leq 0.05$

Next we run a similar firth logit on the subset of the data not connected to the Facebook page (Exchange Relationship). We find that on this subgroup the *Bald-on-The-Record* treatment is the only one which performed significantly better than the control group, this is significant at the ($p \leq .01$). Additionally, the Positive Politeness and the Negative Politeness no longer look to perform significantly better than the control group.

Table 2.11: Logistic Regression Results Exchange Relationship

DV = Conversion	**Coefficient**
Bald on The Record	2.210*** (0.671)
Positive Politeness	1.194 (0.731)
Negative Politeness	.336 (0.829)
Observations	6,742

*** $p \leq .01$, ** $p \leq 0.05$

Permutation Tests

Next, we perform a series of permutation tests to assess the likelihood that our findings are the result of pure chance. To execute this test, we utilize Kaiser and Lacy (2009)'s implementation of a Monte-Carlo Permutated t-test, with 10,000 iterations for all pairwise comparisons that have been reported above. The results comparing each condition in the aggregate sample (ignoring relationship type for the time being) to the *Off-the-Record* condition (our control) are reported in Table 2.12. These tests yield results highly consistent with those reported earlier. Specifically, we observe that the *Bald-on-the-Record* and *Negative-Politeness* strategies outperformed the *Positive-Politeness* strategy at $p \leq .05$ and .01, respectively. The only notable change is that the statistical significance of the difference in conversion between *Positive-Politeness* and *Off-the-Record* decreases to $p = .123$. Broadly, however, the results continue to provide support for the notion that explicit requests outperform implicit requests.

Table 2.12: Permutation Tests: Total Population

Comparison	***p*-value**
Bald on the Record vs Off the Record	.000
Positive Politeness vs Off the Record	.123
Negative Politeness vs Off the Record	.001
Positive Politeness vs Negative Politeness	.047
Positive Politeness vs Bald on the Record	.000
Negative Politeness vs Bald on the Record	.409

Next, we performed the permutation tests on each sub-population, i.e., Communal versus Exchange. These results are presented in Table 2.13. We see that the *Negative-Politeness* condition continued to outperform all other conditions, exhibiting significantly larger outcomes at $p \leq .01$ in each case. This reconfirms our empirical support for Hypothesis 2. We also see that the *Bald-on-the-Record* condition outperforms all other conditions in the Exchange relationship sub-population, yielding significant differences in the outcome at $p \leq .01$ in each case. This reconfirms our empirical support for Hypothesis 3.

Table 2.13: Permutation Tests: Communal vs. Exchange Sub-populations

Comparison	**Relationship**	***p*-value**
Bald on the Record vs Positive Politeness	Communal	.504
Bald on the Record vs Negative Politeness	Communal	.003
Bald on the Record vs Off the Record	Communal	.176
Positive Politeness vs Negative Politeness	Communal	.001
Positive Politeness vs Off the Record	Communal	.488
Negative Politeness vs Off the Record	Communal	.000
Bald on the Record vs Positive Politeness	Exchange	.000
Bald on the Record vs Negative Politeness	Exchange	.000
Bald on the Record vs Off the Record	Exchange	.000
Positive Politeness vs Negative Politeness	Exchange	.180
Positive Politeness vs Off the Record	Exchange	.107
Negative Politeness vs Off the Record	Exchange	.763

Randomization Checks

Randomization was done via Facebook A/B split testing feature. The data reported by the platform is anonymized so that information pertaining to characteristics of the population is limited. However, we were able to execute some randomization checks to ensure balance on their impression device and geographic region.

First we run a Multinominal Logit on whether the randomization was balanced across device types. We weighted the regression based on reach, as Facebook's randomization is done at the individual level [6] . Table 2.14 are the results of the device regression. No coefficients were found to be significant at the $p \leq .10$. As follow up, we also ran separate Multinominal Logits for sub groups of those in Communal and Exchange Relationships, no significant coefficients were found.

Table 2.14: Randomization Check (MLOGIT; DV=Politeness Level)

Variable	Bald on the Record	Positive Politeness	Negative Politeness
android tablet	-0.1839 (0.6767)	-1.6094 (1.0989)	-14.1475 (522.9257)
desktop	0.4448 (0.5345)	0.1542 (0.5632)	0.1350 (0.5632)
ipad	0.3271 (0.7689)	0.6931 (0.7125)	-0.0191 (0.8212)
iphone	0.1154(0.1146)	0.1146 (0.0804)	0.0365 (0.1138)
ipod	-16.44 (3794.546)	-16.5071 (3841.178)	-16.5033 (3797.411)
other	14.348 (1280.433)	15.003 (1280.432)	0.0131 (1796.387)
Constant	-0.0394 (0.08875)	1.27e-08 (0.0879)	-0.0191 (0.08746)
Observations		2,613	
Pseudo R^2		0.0027	
Wald Chi^2		19.54 (18)	

Note: The baseline outcome is Off the Record Politeness

Finally, we ran yet another reach weighted Multinominal Logit to determine if there is balance across states, as this was a US based experiment. For both the overall regression, and sub-regressions we found no significant coefficients at the

[6] https://www.facebook.com/business/help/399737743699353

$p \leq .10$.

Although the individual level data provided by the Facebook advertising platform is limited, from the aforementioned randomization check results we have reason to believe that the randomization process succeeded for this experiment.

2.6 General Discussion

We have offered a first consideration of the role of linguistic politeness in determining the efficacy of social media advertisements (shopping requests) in driving online conversions. We theorize that the appropriate level of politeness depends on the nature of a consumer's relationship with the advertising brand. Conducting a randomized field experiment on Facebook, in partnership with a US-based women's clothing retailer, we begin by providing empirical evidence that consumers are more likely to respond to explicit shopping requests. Subsequently, we offer empirical evidence in support of our theory about the conditional nature of politeness in social media advertising, demonstrating that customers who hold a Communal Relationship with the brand, i.e., those who have 'liked' the brand's Facebook page, are most responsive to a polite request strategy, whereas consumers who hold an Exchange Relationship with the brand, i.e., those who have visited the online store but have not engaged on social media, are most responsive to a blunt, impolite request strategy.

The initial finding that explicit requests are more effective than implicit requests, in general, suggests that the vagary and lack of clarity associated with implicit, *Off-the-Record* solicitations leads to a lack of response that outweighs any undesirable outcomes associated with explicit demands to shop, e.g., reactance. Our results suggest that employing a *Bald-on-the-Record* or *Negative Politeness* request strategy yields optimal conversion for social media advertisements.

Similar to Zemack-Rugar et al. (2017), we have found that less assertive shopping requests lead to better outcomes when directed towards customers who are socially proximate to the brand (in a Communal Relationship). Although Zemack-Rugar et al. (2017) found that non-assertive advertisements, analogous to our *Off-the-Record* messaging of "New Styles", performed better when it comes to likeability, our findings indicate that this strategy is completely ineffective at garnering website conversions. This suggests, first and foremost, that it is not appropriate to assume that social media engagement translates directly into compliance. Secondarily, however, it also suggests that politeness strategies need to be tailored with the advertiser's objective in mind, e.g., online conversion versus social media engagement.

Finally, our finding that politeness is desirable when eliciting conversion from consumers who hold a Communal Relationship, yet undesirable when eliciting conversion from consumers who hold an Exchange Relationship, indicates the importance of aligning politeness strategies with the brand-consumer relationship. When engaging with consumers in an Exchange Relationship, consumers are more responsive to requests made in a blunt, direct fashion, perhaps because attempts to signal solidarity or shared group membership are perceived to be ingenuine and superfluous. In contrast, among consumers who *do* have a pre-existing relationship with the brand, blunt, impolite demands are ineffective, perhaps because they are perceived as incongruent with the norms of the Communal Relationship, and thus come off as rude and uncaring.

Our research design bears strong internal validity; we have utilized theory-driven manipulations, validated our treatments through manipulation checks, and randomized their delivery. At the same time, our study has a number of limitations, primarily tied to external validity. On the positive side, our study is implemented in a field setting, with a real-world retailer, engaging with real-world consumers. That said, our retailer operates within a particular industry (retail fashion), in a particular cultural context (the United States), thus it is possible that our results are somehow dependent upon the nature of the product advertised (Schanke 2017) or the norms of the consumer audience. Moreover, the consumer population we consider is gender imbalanced, being largely comprised of females. However, it should be noted that women drive 70-80% of consumer purchases in the United States, suggesting that the latter limitation should not be of particular concern.[7] Nonetheless, future work might look to explore the potential moderating influence of product features, culture or gender in these relationships we have studied here.

Our findings have important implications for marketers and social media managers. Broadly, we have found that using a more explicit shopping request in social media advertisements yields website conversion. Our results imply that, when engaging with members of one's brand community, social media managers and marketers should remain cognizant of the nuances of their social interactions with consumers. At the same time, our findings suggest that it is reasonable to engage in a direct, blunt manner when interacting with consumers who have yet to engage socially with the brand community.

As noted earlier, this work contributes broadly to the digital marketing literature. We are, to our knowledge, the first study to evaluate the use of imperative requests in social media advertisements, and link the low level linguistic features to their efficacy in garnering off platform outcomes. This work also shows that Politeness levels are an important anthropomorphic feature to be used on social media advertisements. Additionally, this study contributes to the ongoing literature on the importance of linguistics in marketing; and is one of the first to link

[7] Forbes: https://www.forbes.com/sites/michelleking/2017/05/24/want-a-piece-of-the-18-trillion-dollar-female-economy-start-with-gender-bias/#3a0983786123

explicit Politeness Strategies, which allows for precision in the levels of assertiveness, to language used by marketers.

2.7 Conclusion

Language features are an important element of social media content, and firm-generated content in particular. Language that is worded too aggressively or that lacks clarity is unlikely to elicit conversions (though such content may be effective in garnering engagement, e.g., attracting 'likes' or 'retweets'). Politeness, a universal construct that governs communication between individuals (Holtgraves 2011, Brown and Levinson 1987), is an important theoretical framework that marketers should consider, and leverage in their interactions with consumers. As technology and platforms seek to move towards natural language processing-based tools, like chatbots and conversational user interfaces, the role of language in customer-brand interactions will only grow. This study offers an important early step toward improving our understanding of the role of language in online marketing. It is our hope that this work will spur future attention to this line of inquiry, as numerous open questions remain in this space.

Chapter 3

Estimating the Impact of‘Humanizing’ Customer Service Chatbots

3.1 Introduction

Researchers, the general public and organizations alike have become enamored with Artificial Intelligence (AI). With recent breakthroughs in the field, coupled with changes in public perception and advances in hardware, society has seen AI technologies move to the main stage. Organizations are looking to capitalize by putting these technologies into practice to both capture value, and to hedge against the possibility of disruption.[1] AI technologies have seen widespread implementation in a variety of domains, from fraud detection, to image recognition, voice recognition and natural language processing (Dale 2016). Gartner predicts that 2.3 million AI-related jobs will be created by the year 2020.[2]

Although media and public interest have caused AI to reach what Gartner refers to as a state of “inflated expectations”, there is clear value in these technologies, if they are used appropriately and expectations are managed. One prominent example of an AI-based tool that has seen widespread adoption and value creation for firms of all sizes is the text-based ‘chatbot’. Chatbots are autonomous software agents that support text-based exchanges with human users, drawing on tools and techniques from the domain of Natural Language Processing. Chatbots have the potential to automate basic, repeatable, standardized customer service interactions, relieving the need for those interactions to be handled by human employees.[3] Recognizing the potential of these sorts of AI-based autonomous agents, firms are adopting them at an extremely rapid pace. Google Search Trends indicates that interest in chatbots has grown by an order of magnitude in the last two years (see Figure 3.1), and industry estimates forecast that, by 2020, conversations with autonomous agents will be more common for the average individual

[1] https://www.forbes.com/sites/joemckendrick/2018/01/25/artificial-intelligence-isnt-killing-jobs-its-killing-business-models/#64e5388a5ea0

[2] https://www.gartner.com/newsroom/id/3837763

[3] https://chatbotsmagazine.com/chatbots-vs-apps-the-final-frontier-a0df10861c48

than conversations with a spouse.[4]

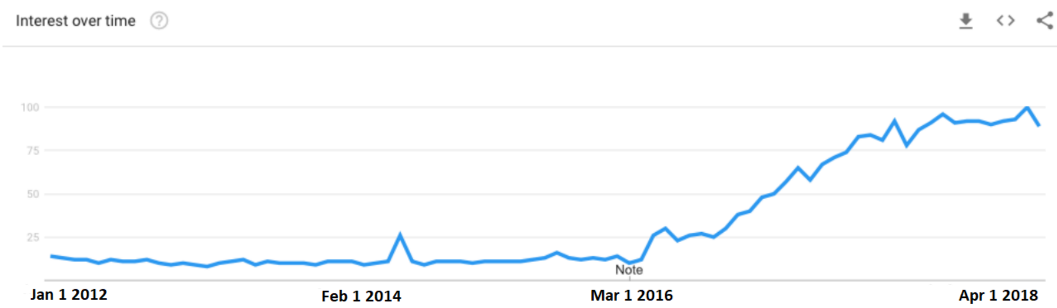

Figure 3.1: Google Trends Global Interest in the Term 'Chatbot'

The anticipated volume of customer interactions these digital agents will be expected to handle suggests that chatbots will soon become the main point of customer contact for many retail organizations. Organizations therefore need to be careful in their design and deployment of these technology artifacts, to ensure that the experience that customers have is both effective and enjoyable. While many features warrant attention, one particularly important aspect to consider is the extent to which autonomous agents (and specifically chatbots) are designed with social interaction, and specifically anthropomorphism, in mind (Wilson et al. 2017).

Though anthropomorphism touches several academic disciplines, it can best be described as the attribution of human-like qualities to non-human entities like machines, animals and other objects (Duffy 2003). This phenomenon is a common

[4] https://www.gartner.com/smarterwithgartner/gartner-predicts-a-virtual-world-of-exponential-change/

occurrence when individuals interact with technology that possess certain elements associated with human-to-human interaction, like eye gaze (Kiesler et al. 1984), facial expressions (Kiesler et al. 2008) and conversational turn-taking (Cassell and Bickmore 2000). How individuals humanize technology has been an important topic of inquiry in both Human Computer Interaction (HCI) and Human Robot Interaction (HRI) literature for decades. In some cases, making technology more human-like has proven to be beneficial, increasing user trust and satisfaction with the interface. However, in other cases, adding human-like social cues has led to negative consequences, such as social anxiety (Sproull et al. 1996) and reduced cooperation (Kiesler et al. 1996). As we articulate in our review of prior literature in later sections, a common feature of much of the prior work in this space is the inconsistency of the relationship between anthropomorphism and desirable user outcomes. This inconsistency speaks to the myriad contextual factors that can shape the relationship. With that in mind, in this work, we seek to understand the impact of integrating anthropomorphic features into AI-enabled autonomous customer service agents, i.e., chatbots, particularly within a retail environment. Specifically, we seek to empirically evaluate the effects of anthropomorphism on transaction conversion. Further, we explore the impact of anthropomorphism on consumer offer sensitivity, informed by prior work in the HCI literature which has drawn a connection from consumer perceptions of anthropomorphism to customer perceptions of fairness and trust. Formally, we evaluate the following two research questions:

- ***RQ1: How and to what degree does customer transaction probability depend on the anthropomorphism of AI-enabled automated customer service agents (chatbots)?***

- ***RQ2: To what degree does customer offer sensitivity vary with the anthropomorphism of AI-enabled automated customer service agents (chatbots)?***

We examine these questions via a field experiment, conducted in partnership with a dual channel clothing retailer based in the United States. Our retail partner has historically operated a used-clothing buy-back program through a web-based form, and employee conversations with customers over email and Facebook messenger. In the prior process, a customer would describe the clothes, obtain a offer estimate from an employee, provide mailing address info and print a shipping label, before sending the clothes to the retailer for final evaluation and payment. We insert ourselves into this process, automating the customer interactions with a Facebook Messenger chatbot, which is integrated with the retailer's Facebook business page. In implementing the chatbot, we integrate a framework that enables us to randomly assign customers into various treatment conditions, such that customers ultimately converse with a chatbot that bears a randomly assigned set of anthropomorphic features. This randomized design allows us to experimentally evaluate the causal relationship between the degree of a chatbot's anthropomorphism and the customer's probability of completing the buy-back process. Moreover, we simultaneously introduce random variation into the cash offer each customer receives, which further enables us to assess the moderating effect of chatbot anthropomorphism on customers' offer sensitivity.

We arrive at two notable findings. First, we find that incorporating anthropomorphism into autonomous customer service chatbots increases conversion rates. Second, we show that, in the presence of a sufficiently large degree of anthropomorphism (3 treatments), customers become more offer sensitive. This latter finding indicates that, as a chatbot becomes more human-like, consumers begin to scrutinize offers. This might occur because offers made by humans are more likely to be perceived as potentially opportunistic (price gouging) or inconsistent (noisy) by consumers, compared to computer-generated offers.

Our study contributes to a number of different streams of literature. First, we contribute to the literature in Information Systems by exploring the design and efficacy of an increasingly prevalent form of information system, the customer service chatbot. In so doing, we build on an extensive literature in HCI related to anthropomorphism by evaluating these features in a field setting. Second, we contribute to the Marketing literature by considering a variety of practical and theoretical issues in the AI-enabled automation of customer service job roles. Building on the work of Wirtz et al. (2018), we empirically evaluate anthropomorphism, a "critical design attribute" of service robots, demonstrating its value in customer service settings. Third, our work contributes to the burgeoning literature on individual's reactions to algorithmic forecasts and estimates (Kleinberg et al. 2017, Dietvorst et al. 2015, 2018, Tambe et al. 2019), and highlight how anthropomorphism could play a role. Finally, and more broadly, our work contributes to the literature on Intelligence Augmentation, or IA (Jain et al. 2018). In particular, our study demonstrates the potential to augment artificially intelligent agents with human-like social intelligence (Wang et al. 2007a). Whereas the literature on IA to date has primarily focused on the possible applications of technology to augment human decision-making abilities, our work highlights opportunities for the reverse; that incorporating human-like behavior and decision-making into autonomous agents can amplify their performance and efficacy as well.

3.2 Literature Review

3.2.1 Anthropomorphism & AI

Scholars of computer science and engineering have dedicated a great deal of attention to the efficient performance of AI-based systems, with an eye toward operational performance. However, when it comes to the automation of job roles or processes that involve human touch-points, social factors are likely to play a particularly prominent role as well. Fortunately, designing autonomous agents to account for social factors has been a focal subject in the Human-Computer Interaction literature for many decades.

A central component in research on the effective design of autonomous agents has been the role of anthropomorphism. Anthropomorphism is a concept that touches several fields of study: psychology (Heider and Simmel 1944, Malle and Pearce 2001, Barrett and Keil 1996), marketing (Aaker 1997), computer science (Duffy 2003, Kiesler et al. 2008) and religion (Guthrie 1995). Although definitions within these fields vary slightly, anthropomorphism, at broad scope, is the attribution of human-like qualities to non-human entities like machines, animals and other objects (Duffy 2003). This attribution is generally the product of humans seeking to explain the actions and behaviors of non-human objects and beings in a way that they understand (Duffy 2003). Although assigning human-like qualities is a very common occurrence that pervades several disciplines, this phenomenon is viewed by several scientific disciplines like biology and psychology as a nuisance that confounds causal mechanisms and hampers scientific inquiry (Kennedy 2003).

While some disciplines view anthropomorphism as a hindrance, others, like HCI, view anthropomorphism as an inevitability that should be accounted for and acknowledged when designing the interface (Caporael 1986). A popular paradigm used in HCI is known as 'Computers Are Social Actors', or CASA, which suggests that people, when presented with technology that contains features like dialogue and turn taking, identify those pieces of technology as a social actor (Moon 2000, Nass and Lee 2001, Nass et al. 1994). It is this conceptualization of digital agents as social actors, that interface designers can apply theories from social sciences, which govern human to human interaction like politeness (Nass et al. 1994) and reciprocity (Moon 2000), and effectively carry these over to human machine interactions (Nass et al. 1994). As such, designers can strategically utilize social cues like small talk, greetings, and transitions to influence user trust with the interface and elicit specific behaviors like self-disclosure (Cassell and Bickmore 2000) and persuasion (Xu and Lombard 2017).

Although anthropomorphic social cues can help designers create a more effective user interface, these features can also lead to unintended negative consequences. More specifically, Ben Shneiderman, a critic of the use of anthropomorphic social cues in the technology interface (Don et al. 1992), contends that designers do not fundamentally understand the way users will perceive and interpret social cues. This lack of understanding can lead to unintended outcomes, namely undesirable perceptions of anthropomorphism (Duffy 2003). As a result, incorporating even minor social cues in an ad-hoc (and ill considered) manner may lead to user disappointment, when the human-like agent falls short of user expectations (Duffy 2003, Nass and Moon 2000). A delicate balance thus needs to be struck when it comes to the incorporation of social cues in chatbots. Accordingly, it should come as no surprise that so many chatbots on Facebook's messenger platform today are incapable of fulfilling the basic requirements of users.[5]

We seek to evaluate the effects of introducing anthropomorphism in chatbots via the three commonly used social cues: social presence, communicative delay, and humor. We will explore how user (customer) exposure to greater levels of anthropomorphism in a chatbot, i.e., greater numbers of features, influence transaction outcomes in a live customer service interaction, as well as any associated shifts in customer offer sensitivity. We discuss the three anthropomorphic features

[5] https://www.fool.com/investing/2017/02/28/facebook-incs-chatbots-hit-a-70-failure-rate.aspx

below, referencing relevant literature for each.[6]

Social Presence: A commonly discussed element in papers related to conversational agents is social presence (Sah and Peng 2015, Verhagen et al. 2014, Araujo 2018). In this technological context, adding social presence means to add "sensitive human contact" (Verhagen et al. 2014). In interacting with a chatbot users have opportunities to make social presence attributions at the beginning (Araujo 2018, Holtgraves et al. 2007), middle (Sah and Peng 2015, Holtgraves et al. 2007) and end (Araujo 2018) of the conversation.

This social presence can prove to be a double edged sword for practitioners. The more socially-present the interactions are, the more engaging the interface; however, the more human-like the interface the higher expectations that the user has of the machine's communicative prowess (Mone 2016, Nowak and Biocca 2003). With this, designers of chatbots make a very important decision of how their conversational agent is perceived in the beginning of the interaction with a greeting (Araujo 2018, Gefen and Straub 2003). For example, a designer can either greet the user, by introducing itself with a real human name, or level

[6] We opt to implement the intensity of anthropomorphism via introducing combinations of treatments, rather than manipulating the level of one treatment, because this enables us to abstract away from any specific cue, to infer effects from anthropomorphism more broadly. In our robustness checks section, we explore the pattern of effects that emerges when we estimate the effect of different combinations of specific cues. There, we demonstrate a pattern consistent with the idea that each cue has a directionally consistent effect on conversion, indicating that our abstraction away from particular cues to anthropomorphism more broadly is justified.

expectations of communicative capability by using a generic machine-like name. By setting the tone with a human name the designer could elicit an anthropomorphic response to the chatbot leading to a more engaging customer experience. Alternatively, in giving the chatbot a human name, the designer could enforce unattainable human expectations on the chatbot, which could lead to frustration later in the experience.

In addition to the greeting, designers can influence anthropomorphic perceptions through the language choices they make in the conversation. For example, using more polite (Fussell et al. 2008), informal (Araujo 2018, Holtgraves et al. 2007) or social (Verhagen et al. 2014) language can help induce anthropomorphic perceptions and also perceptions of social presence. Slight differences in agent language have shown to greatly impact a chatbot's perceived personality (Holtgraves et al. 2007). It is with these linguistic features that designers help to enforce a feeling of social presence and further promote anthropomorphism in their chatbot.

Another method HCI designers use to achieve anthropomorphic attributions towards their machines is through physical social cues (Goetz et al. 2003, Fussell et al. 2008). Unlike embodied conversational agents, chatbots rely solely on text based computer mediated communication to communicate and cannot show physical non-verbal cues like facial expressions or gaze (Kiesler et al. 1984). In computer mediated communication, when these typical face to face social cues are not present, communicators shift focus to alternative cues available and make social interpretations (Walther and Tidwell 1995, Walther 1992). This theory is known as Social Information Processing (SIP), typically this manifests itself in chronemic cues like timestamps (Walther and Tidwell 1995, Liebman and Gergle 2016). Due to the disembodied nature of chatbots that exist on messaging platforms like Facebook Messenger, Kik or Telegram, designers only have a couple of chronemic social cues at their disposal to enforce feelings of a real socially present human. These would include: read receipts and ellipses during typing messages. Although, these two features are common place when two humans are talking via Facebook messenger, these cues are not required from a chatbot as it neither types nor reads.

Although, these anthropomorphic perceptions could lead to the higher amounts of sociability between the chatbot and the customer, these deviations from a more task oriented style could lead to more difficulty and time for users to complete a self-service task. Additionally, it could also over promise the communicative prowess of the agent on the other end of the conversation. This could be counterproductive as users of self-service technologies do so because they are convenient, quick and a means to circumvent interacting with service individuals (Meuter et al. 2000). As such, there is a potential that these communicative features could lead to one of two outcomes. The first is that, the more anthropomorphic the chatbot becomes the more a customer is willing to engage with the artifact. This prolonged interaction would eventually lead to a resolution of the issue, and save labor costs for the company. Alternatively, these anthropomorphic additions to the chatbot obfuscate task oriented nature of the typical self-service interaction, and could lead to frustration and dissatisfaction as the features add overhead to the experience and also mislead the user about the chatbot's communicative prowess.

Communication Delays: In addition to language communication features, another social cue employed by both researchers and practitioners is delay (Holtgraves and Han 2007, Crozier 2017, Gnewuch et al. 2018). From one perspective, delays could be interpreted as the chatbot not working as expected. However, when implemented correctly, slight delays that are dynamic to the amount of text can dictate levels of persuasion (Moon 1999) and chatbot personality perceptions (Holtgraves and Han 2007). At face value, this anthropomorphic effect of delays seems somewhat intuitive as humans do not read and respond to messages sent through text based mediums instantaneously.

Although these slight delays may lead to more anthropomorphic perceptions of the chatbot, they may also interrupt the service quality associated with the experience (Taylor 1994, Meuter et al. 2000). Thus delays in sending messages could lead to two different outcomes in a customer service interaction. If the anthropomorphic features of the interface lead to higher levels of trust in the interface, then potentially these slight delays would enhance the user experience and lead to higher levels of satisfaction with the experience. In contrast, delays can be viewed as an element that impedes the service encounter and prevents the customer from accomplishing the self service task.

Humor: In the fields of socio-linguistics and pragmatics, humor has been shown to introduce feelings of common ground between two communicating social actors (Holtgraves 2011, Brown and Levinson 1987). Similar to human to human interactions, humor can be an effective way to personify systems, and create a more engaging interaction (Niculescu et al. 2013, Morkes et al. 1999). Additionally, humor in task oriented communications has been shown to increase individuals satisfaction with the task (Morkes et al. 1999).

Although humor may be beneficial, it does appear that there is some nuance required in implementing humor. For instance in the medical field, humor helps improve reassurance for patients, but only in the correct context (Francis et al. 1999). This also has been shown in human and robot interaction, where robots with a more playful personality gains more compliance from humans in a non-serious task, and more serious robots perform better in serious task (Goetz et al. 2003). Similarly, humor in both business and customer service interactions requires a more nuanced approach (Malone 1980, Dolen et al. 2008). More specifically, Dolen et al. (2008) find that while humor in an electronic service encounter can help in some situations in which the process is to their liking, but when the process is not to their liking additions of humor exacerbates the negative feelings associated with the service experience. With this nuance of humor, in a customer service interaction, it is unclear whether humor will increase the satisfaction for users engaging with the chatbot or whether it will hinder the overall experience.

Humans and Algorithmic Decision Making. Several emerging studies in Human Resources (Tambe et al. 2019), Economics (Kleinberg et al. 2017) and Psychology (Dietvorst et al. 2015, 2018, Logg et al. 2019) have investigated how humans respond to algorithmic outcomes. Dietvorst et al. (2015) find that in general humans are averse to forecasts made by an algorithm, even when they outperform their less accurate human counter-parts. Dietvorst et al. (2018) further this line of inquiry and find that algorithmic aversion can be reduced when individuals have the ability to manipulate and make adjustments to the algorithm. Similarly, Tambe et al. (2019) theorize that employees will be less accepting of algorithmically determined shift decisions than those determined by a supervisor as they could potentially feel less involved in the decision. Interestingly, Tambe et al. (2019), further discusses an anecdote from Uber, describing that individuals negatively respond to surge pricing when they believe it is set by an algorithm.

Contrasting these findings Logg et al. (2019) find that individuals can be appreciative of algorithmic judgements in numeric forecasts and recommendations for dating and music, as opposed to those made by humans. In addition, Logg et al. (2019) find, similar to Dietvorst et al. (2018), that individuals prefer their own judgements over that of an algorithm. As this aforementioned research indicates, how individuals react to algorithmic outcomes is very dependent on context and human involvement.

Behavioral Economics has sought to understand how individuals reason through offers. One classic example is the *Ultimatum Game*, (Gurth et al. 1982). In this game, a proposer makes an offer of money, and the offer receiver is to accept or reject the offer. The rational expectation is that the proposer is to make a small offer, and the recipient should accept the offer, regardless of its fairness, because this is the utility maximizing response, i.e., take what you can get (Gurth et al. 1982). A fairly robust experimental finding, however, is that offers of 20% of the total funds available are rejected 50% of the time (Sanfey et al. 2003), because of perceived injustice or a lack of fairness.

Previous research has found that human players tend to reject unfair offers less when the actor making an offer is perceived as lacking intentionality, e.g., a computer, rather than a human. For example, Sanfey et al. (2003), Moretti and Pellegrino (2010) report that recipient rejection rates for relatively low offers increase when the offer is made by a human, versus when the offer is made by a computer (notably, a computer that is totally absent of anthropomorphic features). These authors argue that this occurs because human proposers are more likely to induce recipient emotions, such as disgust (Moretti and Pellegrino 2010).

However, other work has documented contradictory evidence. Torta et al. (2013) found that individuals rejected computer generated offers in the Ultimatum Game more frequently than offers made by humans. Torta et al. (2013) theorize that this occurs because human actors have an easier time processing offers from other humans, but face some difficulty deciding how to respond to offers from computers. For example, the willingness to reject an offer may depend on the manifestation or conformity to social norms and etiquette. Thus, whereas a human actor may have no qualms about rejecting an offer from a non-human actor, off hand, social norms might dictate that the human be courteous and considerate when interacting with another human, imposing a sort of social friction on rejection.

More generally, the HCI literature has found that humans respond more socially when computer-based agents are more anthropomorphic (Kiesler et al. 1996, Nass et al. 1994). As one specific example, Kiesler et al. (1996) found that human participants presented with a Prisoner's Dilemma game tended to respond socially to 'humanized' computer actors, in a manner similar to the response they would exhibit with a true human partner. These findings further the notion that a potentially important element leading to offer receivers acceptance or rejection of offers is the level of anthropomorphism of the automated proposer.

As there is ample evidence to support the benefits and detriments of including anthropomorphism in customer service chatbots, we take on this study and look to its data to help us reach a conclusion.

3.3 Study Context

As described above, we conducted our field experiment in partnership with a dual channel clothing retailer based in the United States, similar to other businesses like Plato's Closet and Clothes Mentor. This retailer buys and sells women's used clothing, both online and through three brick and mortar locations in Iowa and Minnesota. We replaced the retailer's prior, manual clothing buy-back process with an AI-enabled chatbot. The process we automate was previously managed via web-form and email exchanges, or done in person at a store. We developed the chatbot using Google's Conversational AI Platform, DialogFlow, incorporating Python-based customizations. DialogFlow enables the automated processing and generation of conversational prompts and utterances in exchanges employing natural language. The Python customizations were incorporated to implement required business rules and logic, as well as to manage the conversational flow (e.g., if customer says this, do that). The chatbot was integrated with the retailer's Facebook business page, as part of the retailer's Facebook messenger profile. The retailer's Facebook page has approximately 44,000 followers.

The chatbot is designed to interact with customers who are interested in selling their used clothing to the retailer. The overall conversational interaction model has three major steps. First, the chatbot begins by requesting information on the number and types of clothing that the customer wishes to sell. Then, the chatbot provides an estimated cash offer, indicating the expected value that the retailer would be willing to pay for the clothing described. If the customer accepts the offer, the chatbot then requests additional personal details that are required to complete the transaction, including a mailing address, full legal name, and phone number. Based on this information, a shipping label is generated, which the customer can print and use to send their clothes to the retailer.

3.4 Methods

3.4.1 Experiment Design

To causally identify the impact of the aforementioned anthropomorphic features on transaction outcomes, we implement three independently randomized treatments, one associated with each of three anthropomorphic features. When a customer initiates a conversation with the chatbot for the first time, he or she is randomized into receiving zero, one, two or all three of the anthropomorphic features, in random combinations. We describe the implementation of each treatment, below. Note that by independently randomizing each anthropomorphic feature, we ensure that there is no association between the number of features a customer receives, and which features a customer receives. Our randomization is performed on a between subjects basis. If a single customer revisits our chatbot and initiates additional conversations with our chatbot, we exclude any such subsequent observations from our analysis.

It is worth highlighting that our focus is not on any one of the anthropomorphic treatments, but rather on the number of treatments a subject receives. Our objective in delivering varied numbers of treatments is to causally shift a subject's perception of anthropomorphism in the chatbot interaction. Conceptually, this approach is analogous to the notion of *Combination Therapy* or *Polytherapy* in medicine, which refers to treating a single disease with multiple types of interventions, in concert (e.g., Möttönen et al. 1999). We opt for this approach, rather than attempting to manipulate the intensity of a given anthropomorphic feature by shifting its level, for two reasons. First, it is not altogether clear how dosage manipulations could be achieved with each of the treatments, e.g., it is not altogether clear what would constitute more versus less humor. Second, the perception that one is certainly interfacing with a human actor is unlikely to be achieved through a single manipulation, even in a text-based setting. A chatbot that responds instantaneously, yet also drops a joke into the conversation, may be perceived as having some human traits. However, it is unlikely that simply adding more jokes into the exchange will achieve further improvements. Thus, it is reasonable to assume that anthropomorphism depends a great deal on delivering a sufficient constellation of anthropomorphic features as part of the exchange.[7]

Additionally, for all customers, we introduce random variation into the cash offer. In the original buy-back process, the retailer would calculate an initial cash offer based on a fixed amount of $3.50 per clothing item. We randomly perturbed the offer around the fixed baseline offer for each customer, drawing from a random normal distribution with mean 0 and variance 0.5. That is, our offer perturbations

[7] We offer later analyses, namely manipulation checks, which indicate that perceived anthropomorphism is increasing in the number of treatments received, providing support for our argued mechanism.

were implemented by taking the $3.50 baseline offer previously employed by the retailer, and adding a random value drawn from this normal distribution. Drawing from a normal distribution allowed us to accommodate concerns on the part of the retail partner that cash offers would be 'too extreme' in either direction, creating customer experience issues on the one hand and economic losses for the retailer on the other hand.

Social Presence. To operationalize anthropomorphic social presence, we do so through a combination of a name, linguistic features and social cues related to reading and authoring messages. We thus adopt a methodology similar to that of Araujo (2018). More specifically, in this treatment, we first give the chatbot a randomly drawn human name from the 1990 census, which the chatbot uses to introduce itself at the outset of the conversation. Second, like Araujo (2018), the chatbot employs relatively informal, casual language (as opposed to more formal, professional language). An example of the initial greeting manipulation can be found in the table below.

Table 3.1: Social Presence Manipulation

Condition	**Message**
0	*"Hello I am an automated service bot here to assist with shipping previously used maternity clothing for money."*
1	*Hi I'm Teddy here to help you with shipping previously loved maternity clothes for* $

In the human-like condition, users will also see the cues typically associated with messages exchanged between humans. On the Facebook Messenger platform, these cues include both read receipts when a message is sent to the chatbot, as well as the display of a cue indicating that the chatbot is typing a message. An example of the typing feature can be seen in Figure 3.2 and read receipts in Figure 3.3.

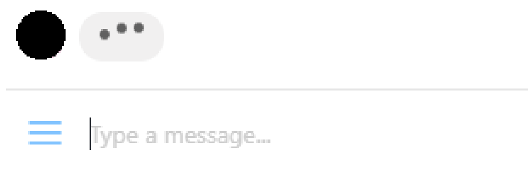

Figure 3.2: Typing Feature

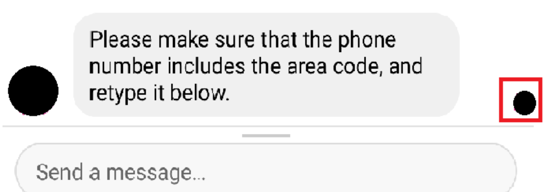

Figure 3.3: Read Receipt is Shown as Small Profile Image on Right

In conditions where these cues are not present, the user sees simply the white messenger background without the read receipts or typing features.

Communication Delays. Similar to Moon (1999) and Holtgraves and Han (2007), we implement a dynamic delay of 70 words per minute. This is within the range of those that type professionally [8] . In the non-human-like condition, users

[8] https://www.livechatinc.com/typing-speed-test/#/

will experience instant responses.

Humor. To operationalize the humor construct we insert a random joke drawn from an approved list of 4 jokes. These jokes were deemed to be inoffensive, and suitable for any age. The random jokes are added into the dialogue, right before the customer receives the estimate for the clothes they will be selling to the retailer. In conditions that do not have humor present, the customer is asked if they will wait a moment while the chatbot totals up their estimate, and a 5 second long pause ensues. This interaction is depicted in Figure 3.4. A brief summary of all manipulations can be found in Table 3.2.

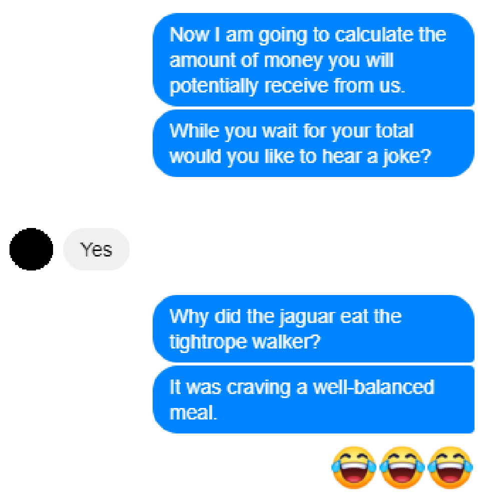

Figure 3.4: Joke Example

Table 3.2: Chatbot Features

Feature	Description
Social Presence	*Human Name, Informal Language, Typing Cues*
Delay	*Dynamically typed 70 WPM delay*
Humor	*Randomly selected joke before estimate*

3.4.2 Empirical Specification, Variables & Data

In our analyses, we are interested in understanding the effect of increasing 'humanization' of the chatbot on i) the probability of conversion, and ii) the moderating effect on the relationship between randomly varied offer amount, and conversion. Accordingly, our primary outcome variable of interest is a binary indicator of conversion. Our independent variables include a series of dummy variables reflecting different levels of the number of anthropomorphic treatments a subject received, *Treatment_Count*, as well as a measure reflecting our offer perturbation, *Cash Offer*, which we mean-center for the sake of simplicity.

We first estimate a series of Linear Probability Models (LPMs), regressing conversion on our treatment count dummies and our offer deviation measure, to understand their direct effects. Subsequently, we interact the dummies and the offer measure, to understand the moderating effects of interest, i.e., how increasing anthropomorphism moderates offer sensitivity. Our final cash offer sensitivity model is reflected below in Equation 3.1, where subjects are indexed by i.

$$\text{Convert}_i = \alpha + \beta_1 \cdot 1\ Treatment_i + \beta_2 \cdot 2\ Treatments_i + \beta_3 \cdot 3\ Treatments_i + \delta \cdot Cash\ Offer_i + \gamma_1 \cdot 1\ Treatment_i \cdot Cash\ Offer_i + \gamma_2 \cdot 2\ Treatments_i \cdot Cash\ Offer_i + \gamma_3 \cdot 3\ Treatments_i \cdot Cash\ Offer_i + \epsilon_i \tag{3.1}$$

Our experiment includes 323 subjects who initiated a conversation with our chatbot between November 16^{th} and December 31^{st} of 2018. We present the descriptive statistics for our variables in Table 3.3. As can be seen, approximately 8.36% converted, meaning they completed the buy-back procedure and obtained a shipping label to send their clothes to the retailer. We also observe that the average user received 1.5 anthropomorphism treatments. Figure 3.5 depicts the distribution of randomized per item offers that were assigned to subjects. As explained earlier, the distribution of offer deviations is normal.

Table 3.3: Descriptive Statistics

Variable	**Mean**	**St. Dev.**	**Min**	**Max**
Social Presence	0.56	0.50	0.00	1.00
Delay	0.48	0.51	0.00	1.00
Humor	0.46	0.50	0.00	1.00
Treatment Count	1.50	0.89	0.00	3.00
Cash Offer	-0.02	0.68	-1.82	1.43
Conversion	.0836	.2772	0	1

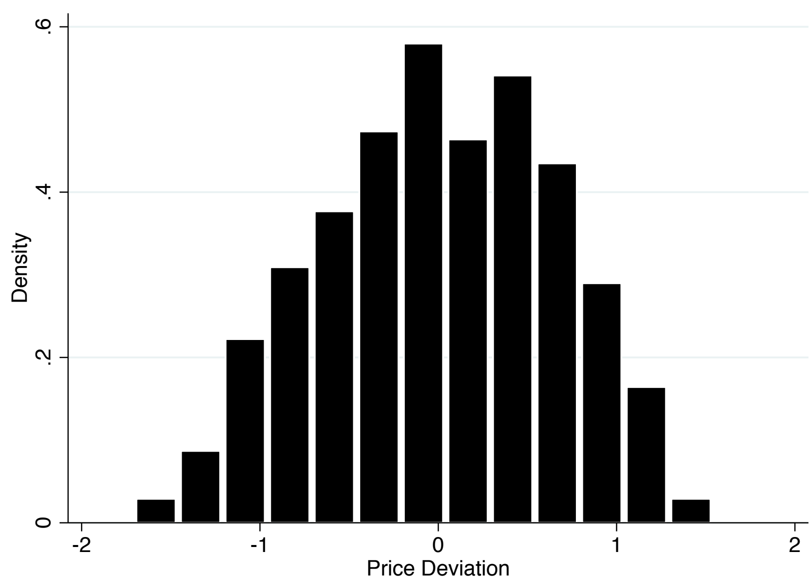

Figure 3.5: Distribution of Per Item Offer Deviation

3.5 Results

We begin by estimating a linear probability model, incorporating only the main effects of each variable. We then progress to incorporating interactions, to recover any effect of cash offer increases on conversion outcomes under alternative levels of anthropomorphism.

Considering the results in Table 3.4, in Column 1, the constant term indicates that the baseline rate of conversion in the control condition (no anthropomorphic treatments) is approximately 2.6%. We observe positive coefficients associated with all other variables in the model. Specifically, we observe that a single anthropomorphic treatment is associated with a 6.7% increase in the probability of conversion ($p < 0.10$), relative to control; a pair of treatments is associated with a 5.0% increase in the probability of conversion (though the result is not statistically significant relative to a null hypothesis of 0); and the receipt of all three treatments in tandem is associated with a 10.8% increase in the probability of conversion ($p < 0.05$). Although the coefficient on cash offer is positive as we expect (given this is a cash offer made to the customer, not a cash offer charged), the coefficient is not statistically significant. That said, the estimate indicates that a \$1.00 increase in the cash offer is associated, on average, with a 2.7% increase in the probability of conversion.

Table 3.4: Treatment Count Model (LPM)

Variable	DV = Convert	DV = Convert
1 *Treatment*	0.067* (0.036)	0.076** (0.032)
2 *Treatments*	0.050 (0.034)	0.060** (0.030)
3 *Treatments*	0.108** (0.055)	0.109** (0.049)
1 $Treatment \cdot Cash\ Offer$	—	0.052 (0.064)
2 $Treatments \cdot Cash\ Offer$	—	0.086 (0.069)
3 $Treatments \cdot Cash\ Offer$	—	0.211** (0.087)
$Cash\ Offer$	0.027 (0.022)	−0.058 (0.057)
Intercept	0.026 (0.023)	0.017 (0.017)
Observations	323	323
R^2	0.016	0.037
F	1.60 $(4, 319)$	3.85*** $(7, 316)$

Note: Robust SEs; ** $p < 0.05$, * $p < 0.10$.

Next, considering the interaction model in Column 2, the main effects associated with the intensity of anthropomorphism remain quite consistent, except that all three estimates are now statistically significant at commonly accepted thresholds (when our cash offer manipulation is 0). Additionally, considering the cash offer interactions, we see that all coefficients are positive and increasing in the number of treatments. Of particular note, we observe that the cash offer manipulation has a statistically significant interaction with the delivery of three anthropomorphic treatments, relative to the delivery of none ($p < 0.05$). This finding indicates that, in the presence of sufficient anthropomorphism, consumers become significantly more offer sensitive.

3.6 Robustness

3.6.1 Estimator Choice & Regression Specification

We begin by considering the robustness of our results to possible concerns of multicollinearity, as well as to our choice of estimator. We report analyses addressing possible concerns of multicollinearity in Appendix A.1, where we provide evidence that this is not a serious concern in our analysis. Subsequently, in Appendix A.2, we explore the robustness of our results to our choice of estimator, namely the Linear Probability Model. There, we demonstrate that our results remain stable under alternative estimator choices.

3.6.2 Replication

We next assessed the replicability of our main finding, that anthropomorphism increases transaction rates, conducting a second, simpler experiment in the same field setting. With this replication, we sought to again address possible concerns that our results somehow derive from aggregating across multiple treatments. With that concern in mind, we sought to evaluate the treatment effect of just a single anthropomorphism treatment, relative to a control condition. This replication thus allowed us to assess whether, given sufficient power, a single anthropomorphism intervention would yield statistically significant estimates of increased conversion. We focused on the social presence treatment in this replication, because it is the intervention that aligns most intuitively with anthropomorphism (Araujo 2018).

The replication was conducted in the same field context. The only distinction in this case is that our experiment was limited to just two conditions: the control condition, in which no anthropomorphism treatment was delivered, and the social presence condition. As before, we assessed the relationship between the treatment and the probability of successful conversion. This experiment was carried out over a 1-month period, from late June to late July of 2019. Recruitment for the replication study was conducted in the same manner, employing Facebook messenger advertisements.

This experiment involved 546 subjects, who were approximately balanced in their assignment to treatment and control; the mean value of our treatment indicator, *Social Presence*, was 0.46. As before, we regressed a binary indicator of transaction conversion onto a treatment dummy, employing a Linear Probability Model. As before, we observe a positive, statistically significant effect on conversion rates with this single, individual treatment. Specifically, social presence features led to an approximate 5% increase in the transaction conversion rate (p = 0.046). Thus, we successfully replicate the main result. Moreover, we conclude that, given sufficient statistical power, we can detect that a single anthropomorphism treatment can translate to tangible benefits for transaction conversion.

3.6.3 Manipulation & Randomization Checks

We performed a manipulation check with 19 volunteers, to ensure that the various treatments were properly experienced by users, and that they had the expected effects on both anthropomorphism level and perceptions of manipulations. To determine if end users indeed experienced the delay and humor treatments, we asked participants to rate their agreement with certain statements, on a scale 1 (Strongly Agree) to 6 (Strongly Disagree). For the humor treatment, the statement was: *The customer service agent was humorous.* For the delay treatment, the statement was: *The customer service agent took a long time to respond.* To analyze the survey responses we used the Mann-Whitney U test (Mann and Whitney 1947). We find that the there is a significant difference between responses that were in the humor and non-humor conditions and the delay and non-delay condition. This is significant at the p $\leq$.01 level.

In addition to running the tests for both the humor and delay manipulations, we also tested whether the delivery of these features in tandem with linguistic

Table 3.5: Results Mann-Whitney Rank Sum Test for Manipulation's Perceptions of Delay & Humor

Condition Comparison	$p - value$	**z**
Humor vs Non-Humor	0.0045	2.842
Delay vs Non-Delay	0.0006	3.417

features led to a higher perception of anthropomorphism. To test this, we used a semantic differential scale, including survey items first introduced by Powers and Kiesler (2006). These survey items are also a component of the Godspeed Questionnaire (Bartneck et al. 2008), a widely used survey in the HCI and Human Robot Interaction literature to measure anthropomorphism (Weiss and Bartneck 2015). The semantic scale ranges from 1 to 6, for five binary word associations: (Fake, Natural), (Machine-like, Human-like), (Unconscious, Conscious), (Artificial, Life-like), (Moving Rigidly, Moving Elegantly). The lower the score, the less anthropomorphic the artifact is perceived to be. Note that we adapted the final word-pair to our textual context, replacing it with (Messages Rigidly, Moving Elegantly). The original scale was developed for use with physical artifacts, i.e., robots, to capture perceptions of movement in physical space; however, because our artifact only exists on the Facebook messenger platform, slight modification was necessary. We averaged the values across the 5 semantic differential scale items to arrive at our final measure.

To determine if the addition of these features leads to higher perceptions of anthropomorphism, we sum the treatment dummies associated with the features: Social Presence, Communication Delays, and Humor, such that we construct a measure capturing the number of treatments a subject receives (which we expect to associate with increasing levels of perceived anthropomorphism). We then perform an Ordinary Least Squares regression of the mean anthropomorphism differential scale response against the count of treatments received. Doing so, we find a statistically significant, positive association ($\beta = 0.619$; $p < .10$). This manipulation check parallels our main analyses, described earlier, in which we explore the relationship between the number of treatments a subject receives, and their conversion response. Conceptually, our approach is analogous to the notion of *Combination Therapy* or *Polytherapy* in medicine, which refers to efforts to tackle a single disease with multiple treatments, in tandem (e.g., Möttönen et al. 1999). Measures similar to that we employ here have been advanced in the medical literature, i.e., based on a summation over treatment interventions received by a patient or subject (Frei et al. 1998). Thus, rather than attempt to manipulate the intensity of anthropomorphism by shifting the levels of any given treatment (it is not altogether clear what would constitute more versus less humor, or greater versus less social presence), we opt for the delivery of more versus fewer treatment options, in combination, to achieve our manipulations.

In addition to these manipulation checks, we also conducted a number of randomization checks, to assess the efficacy of our randomization procedure. Because we randomize in real-time, as subjects arrive, and only have a small set of information describing our subjects available from Facebook, we are limited in the types of randomization checks we are able to perform. As such, one check we can perform is to assess the significance of the association between the number of treatments a subject was assigned and the day on which they entered our sample. To assess this, we perform a Multinomial Logistic Regression of the number of treatments assigned on a vector of day of week indicators. We report the results of this regression in Table 3.6, where all coefficients are statistically insignificant. A similar analysis performed as a ordinal logistic regression also yields null results. This provides some assurance that our randomization procedure was effective.

Table 3.6: Randomization Check (MLOGIT; DV=Treatment Count)

Variable	Treatments = 1	Treatments = 2	Treatments = 3
Tuesday	0.872 (0.696)	0.280 (0.722)	0.118 (0.859)
Wednesday	1.034 (0.689)	1.069 (0.683)	0.929 (0.774)
Thursday	0.178 (0.599)	0.118 (0.596)	-0.352 (0.750)
Friday	0.588 (0.661)	0.057 (0.687)	0.300 (0.778)
Saturday	0.523 (0.630)	0.463 (0.627)	0.405 (0.728)
Sunday	0.187 (0.620)	0.554 (0.598)	-0.442 (0.794)
Constant	0.575 (0.417)	0.636 (0.413)	-0.118 (0.487)
Observations		324	
Pseudo R^2		0.014	
Wald Chi^2		10.94 (18)	

Note: The baseline outcome is 0 Treatments; Robust SEs.

Beyond this assessment of inter-temporal randomization, we also assessed randomization efficacy in two other ways. Specifically, we assessed possible systematic associations between the per-unit cash offer and the treatments a subject was assigned, as well as possible systematic associations between the per-unit cash offer and the number of clothes a subject wished to sell. Each evaluation was conducted via a series of pairwise t-tests, testing for significant differences in pairwise group means. This was done both in terms of treatment count assignments, as well as specific treatment assignments. In all cases, we observe statistically insignificant differences across groups. These results are presented in Appendix A.3.

3.7 Mechanism Exploration

Although we have demonstrated a robust, positive, causal relationship between anthropomorphism features and transaction conversion, it is important to also assess the boundary conditions for our findings, as well as to assess the extent to which anthropomorphism is the primary mechanism behind this relationship. Accordingly, we undertook a variety of secondary analyses and controlled experiments. We first sought to better understand the extent of perceived anthropomorphism associated with our most anthropomorphic chatbot, and how it compared with an obvious benchmark, namely a true human agent. This exercise is important, because it speaks to the potential for further gains, above and beyond the anthropomorphism levels we implemented in this study.

To assess this question, we recruited 54 turkers from Amazon Mechanical Turk and assigned them to either interface with i) our most anthropomorphic chatbot, or ii) a human agent, drawn at random from a pool of four graduate research assistants.[9] These human customer service agents were given a high-level verbal instruction about the information they needed to supply and collect from visitors to complete the buy-back process, including examples of past chatbot interactions.

Each research assistant received a brief training session with one of the authors, and each was observed in a customer service interaction before the experiment was begun to ensure proper understanding of the script. Subsequent to interacting with a customer service agent (either the chatbot or a human), the turkers were asked to respond to a pair of survey items, rating their perceptions of the respective agent's anthropomorphism. To gauge anthropomorphism, we utilized a semantic differential scale, including survey items first introduced by Powers and Kiesler (2006), which ask the subject to rate their interaction on a 1 to 6

[9] The use of multiple human agents is particularly important for this analysis, if we wish our results to be plausibly generalizable. If we were to compare our chatbot against a single human agent, it would be quite difficult to draw conclusions about how the bot might compare to human agents, broadly, versus the particular human agent participating in the study.

scale for five binary word associations: (Fake, Natural), (Machinelike, Human-like), (Unconscious, Conscious), (Artificial, Lifelike), (Messages Rigidly, Messages Elegantly).

The results of this comparison are presented below in Figure 3.6, which depicts group means and 95% confidence intervals. A Mann-Whitney U test indicates that a randomly drawn human agent was perceived to be more anthropomorphic than the fully anthropomorphic chatbot, to a statistically significant degree ($p < 0.05$). The difference on a 6-point scale is 2.97 vs. 3.93, this finding does suggest that there is room to further increase perceived anthropomorphism of our chatbot, and perhaps garner greater benefits for transaction outcomes.

Next, we sought to understand the extent to which our results might derive from our anthropomorphic treatments causing subjects to believe they were truly interfacing with a human agent, versus whether subjects were aware the agent was autonomous and were merely personifying its behavior. Understanding this aspect is important for two reasons. First, there has recently been a push from government regulators to require the disclosure of agents' autonomous nature at the outset of any customer interactions.[10] Accordingly, from a practical perspective, if our results are somehow dependent on the absence of formal disclosure, this

[10] https://www.natlawreview.com/article/get-all-your-bots-row-2018-california-bot-disclosure-law-comes-online-soon

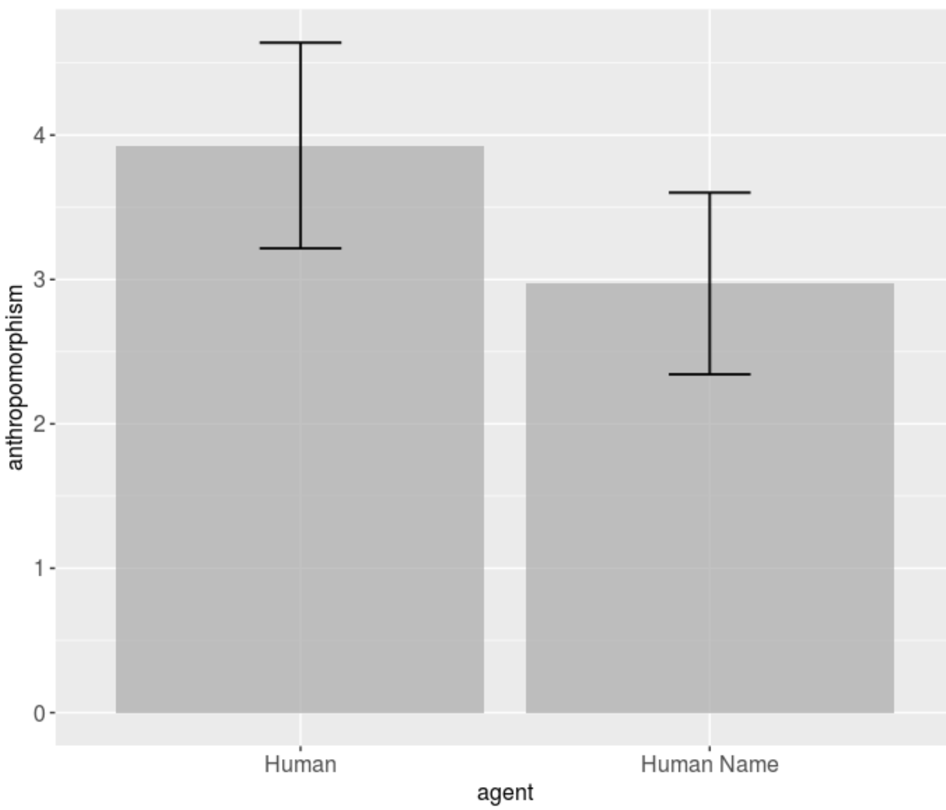

Figure 3.6: Perceived Anthropomorphism - (L) True Human vs. (R) Anthropomorphic Chatbot

would be undesirable, as the value of these findings would be undercut by ongoing regulatory changes in the market. Second, recent work involving voice-based chatbots has reported that a failure to disclose a bot's autonomous nature at the outset of interactions can have detrimental effects on transaction outcomes, if a customer initially believes the agent to be a human, and discovers its autonomous nature only later (Luo et al. 2019).

Our analysis was conducted in a manner similar to the above anthropomorphism bench-marking exercise. Specifically, we recruited 52 turkers to interface with one of two chatbots: i) our fully anthropomorphic chatbot (which lacks explicit disclosure that it is autonomous) and ii) our fully anthropomorphic chatbot, incorporating disclosure. Up-front disclosure was achieved in the latter case by removing the human name and replacing it with the title 'Customer Service Chatbot'. Again, subsequent to these turkers' interactions with their assigned agent, we asked them to respond to survey items. Because we lack objective transaction outcomes in this context, we instead relied upon a proxy response, namely an indication of likeability. For this purpose, we employed adaptations of the survey questions from Mathur and Reichling (2016), obtaining responses to the following prompt: "rate how enjoyable/unpleasant it was interacting with your customer service agent," responding using a sliding scale from -100 to 100. The results are presented below in Figure 3.7, which again depicts group means and

95% confidence intervals.

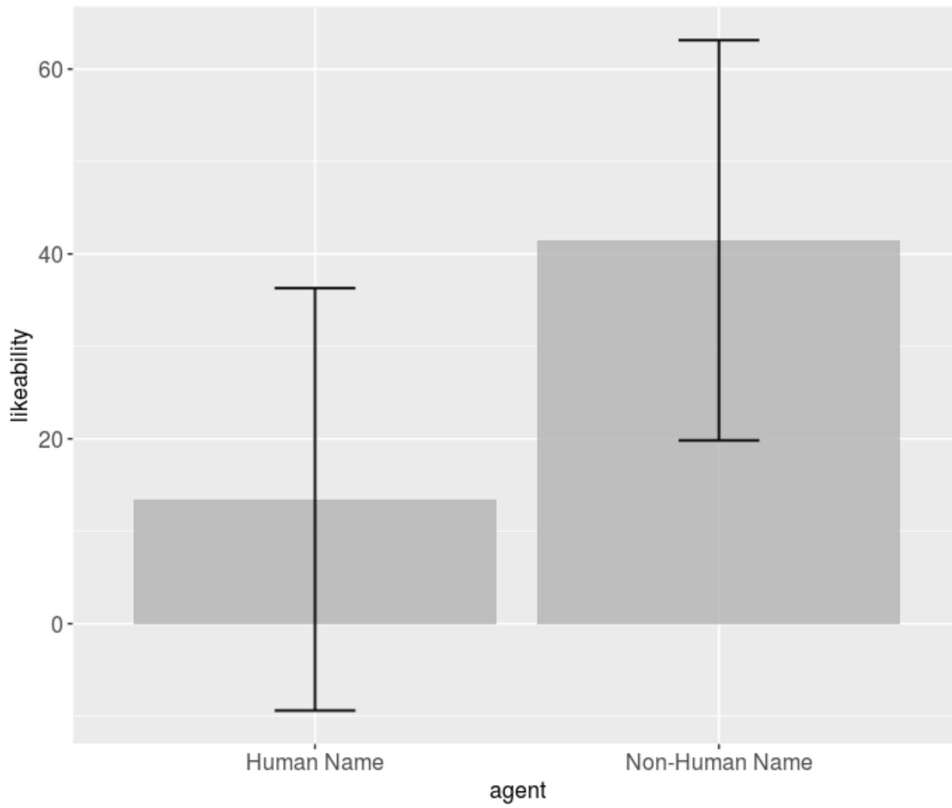

Figure 3.7: Perceived Likeability - (L) Undisclosed Chatbot vs. (R) Disclosed Chatbot

Interestingly, in this case, we find that, counter to expectation, the fully anthropomorphic chatbot without disclosure was perceived to be significantly *less* likeable than the same chatbot incorporating disclosure ($p < 0.10$). Importantly, this finding indicates that the increases in transaction rates are not dependent upon a lack of disclosure that the agent is autonomous. To the contrary, explicit disclosure appears to *improve* customer perceptions. It is plausible that this occurs because, in our context, users can very quickly deduce that the agent is not human, based on its conversational behavior (even without disclosure). Thus, when the chatbot initially presents a human name, this may create an expectation of human interaction, only to be let down shortly thereafter when the customer perceives that responses are automated. What is more, such rapid realization of the chatbot's autonomous nature may lead customers to perceive some attempt at deception. Under this logic, our findings are in fact consistent with those recently reported by Luo et al. (2019), who found that individuals reacted negatively to delayed disclosure of a chatbot's autonomous nature, versus earlier disclosure.

Having evaluated the anthropomorphism of our chatbots relative to human agents, and having considered whether our results are somehow dependent upon a lack of disclosure, we next turned our attention to an exploration of the underlying mechanisms by which anthropomorphism may benefit transaction outcomes. Our earlier offer sensitivity result speaks to this somewhat, in that it suggests that subjects think differently when engaging with an anthropomorphic chatbot. However, we wished to identify concrete evidence of how this differential mindset may benefit transaction outcomes.

One particularly plausible mechanism pertains to humans' trust and willingness to engage in information sharing with autonomous agents. Prior work has observed that a socializing technology can lead to increased persuasion of users

(Holzwarth et al. 2006, Wang et al. 2007b) and can lead to more intimate self-disclosure (Moon 2000). In a customer service interaction, social cues may thus lead to greater comfort with the automated customer service agent, on the human customer's part, which then leads to increased levels of information sharing (Sproull et al. 1996). It is therefore possible that the positive relationship between anthropomorphism and transaction conversion is driven, at least in part, by customers' increased willingness to share sensitive data with the customer service agent that is necessary to complete the transaction.

To explore this possibility, we revisited our original experimental results, considering the treatments' relationship with different information disclosure milestones within the clothing buy-back process. After the offer is seen by a subject, the chatbot proceeds to ask a series of questions to collect contact information that is necessary to complete the transaction. Some of that information is innocuous (i.e., the required dimensions for a shipping box), whereas other information is relatively sensitive (i.e., mailing address, legal name, telephone number). In Table 3.7, we present the results of repeating our main regression using these different milestones as alternative dependent variables.

As we can see from the results, the anthropomorphic treatments begin to have a statistically significant effect as the customer moves further into the process, as the information becomes more sensitive. Although exploratory in nature, these initial results suggest a partial explanation for the effects we see. Certainly, they point to a potentially fruitful area for further inquiry and policy debate around the incorporation of features aimed to achieve anthropomorphism in autonomous, customer-facing agents.

Table 3.7: Information Disclosure Milestones (LPM)

Variable	DV = Box Size	DV = Mailing Address	DV = Legal Name	DV = Phone Number
1 *Treatment*	0.042 (0.0546)	0.063 (0.043)	0.078** (0.036)	0.078** (0.036)
2 *Treatments*	0.061 (0.0558)	0.056 (0.043)	0.071** (0.036)	0.071** (0.036)
3 *Treatments*	0.066 (0.0712)	0.113* (0.064)	0.136** (0.060)	0.136** (0.060)
Intercept	0.093 (0.045)	0.047 (0.032)	0.023 (0.0231)	0.023 (0.0231)
Observations	323	323	323	323
R^2	0.004	0.009	0.015	0.015
F	0.46 $(3, 319)$	1.30 $(3, 319)$	2.88** $(3, 319)$	2.88** $(3, 319)$

Note: Robust SEs; ** $p < 0.05$, * $p < 0.10$.

3.8 Discussion & Conclusion

Our study offers a novel glimpse into how chatbot anthropomorphism, in a real-world customer service setting, influences business outcomes. We explore prior design theory from HCI, which speaks to the consequences of incorporating anthropomorphic features into an autonomous agent, and the implication for various social outcomes, e.g trust. Although there is reason to believe that trust will lead to customer satisfaction, thereby translating to economic benefits for the firm, it is important to recognize that customer trust and satisfaction with a service provider are only two mediating factors that determine transaction outcomes. For example, although a customer may be more trusting of a 'human-like' autonomous agent, they may simultaneously perceive operational inefficiency, and then opt to transact with an alternative provider. Nonetheless, our results are consistent with the notion that anthropomorphic features have a direct, beneficial relationship with transaction outcomes. Our findings are also consistent with prior studies of anthropomorphism's impact upon trust.

Interestingly, we also find that while anthropomorphism influences transaction conversion positively, it also impacts a customer's offer sensitivity. While our context is somewhat unique to retailers, our findings do give reason to believe that high levels of anthropomorphism is not to be incorporated in all customer service chatbots, and its benefits may be dependent on contextual factors. We also find that anthropomorphism, in our context, plays the most important role in sensitive information disclosure. More specifically, we analysed how anthropomorphism influenced conversion of intermediate variables within the buyback process, and found that it plays a bigger role as customers input more personal information. Though preliminary, this highlights that in certain contexts in which firms require information from their customer, high levels of anthropomorphism could be advantageous. In further experiments discussed in the Appendix on Mechanical Turk, we also find that the individual treatment drives likeability of the agent, and this in turn could be driving much of these conversion outcomes.

Another notable finding comes from our follow-up studies involving crowd-workers. We sought to evaluate whether the practice of disclosing the chatbot's autonomous nature would influence user perceptions of likeability (our proxy for customer satisfaction). Ultimately, we found that disclosure (i.e., a chatbot that uses a name like 'Customer Service Chatbot') was *more* likeable than the undisclosed chatbot (employing a human name). As we noted earlier, we believe this occurs because customers quickly come to realize that they are not interacting with a human, even in the absence of explicit disclosure. Whereas disclosure makes this clear immediately, a failure to disclose may thus translate to delayed (and unplanned) disclosure, which customers could interpret as an attempt at deception, or falling short of their expectations (Oliver 1977). This finding once again points to the importance of context, and customer expectations. If customers are operating in an environment where they anticipate engaging with automated customer service agents, their expectations for the exchange may be quite different than alternative settings in which a human agent is expected. Recent research has observed that many consumers have grown more comfortable with the notion of algorithms in their daily lives, going so far as to exhibit 'algorithm appreciation' (Logg et al. 2019). This aspect is important for firms considering the design and implementation of autonomous customer service agents.

Additionally, chatbots represent a means by which firms can ensure consistent performance in their human facing customer service roles. In many customer service jobs, individuals are expected to perform routinized tasks with nearly mechanistic efficiency and perfection. This is difficult because individual workers behave differently from each other, as well as the same individual varies their behavior throughout the day. This standardization of service delivery is both a chief concern among most retailers today,[11] as well as a key reason many firms are considering implementing autonomous customer service agents.[12] As such, a potentially effective compromise, that simulatenously leverages the social intelligence of humans, in tandem with the standardized delivery enabled by autonomous agents, is to imbue chatbots with social intelligence (Wang et al. 2007a). Although current conversational technologies are unlikely to replace the best human customer service agents in the short term, it is plausible that socially intelligent chatbots could lead to improvements in the customer experience if employees exhibit issues with consistency of service delivery and service experience. This observation resonates with the findings of Luo et al. (2019) that autonomous agents may perform better than inexperienced workers in a sales context.

[11] eMarketer - Leading Challenges Facing Retailers: https://www.emarketer.com/chart/229895/leading-business-challenges-facing-in-store-retail-according-us-retailers-may-2019-of-respondents

[12] Drift - 2018 State of Chatbots Report: https://www.drift.com/wp-content/uploads/2018/01/2018-state-of-chatbots-report.pdf

Our research also points to possible opportunities for intelligence augmentation (Jain et al. 2018). First, our work demonstrates that augmenting AI-enabled autonomous agents with human-like social intelligence can increase their performance in customer service settings (Wang et al. 2007a). What is more, our research design suggests a procedure by which firms might leverage autonomous chatbot implementations to experimentally evaluate the most effective patterns of customer interaction, with an eye toward informing the training of human customer service agents. For instance, our experimental results demonstrated that some degree of humor (discussed in Appendix A.4) can lead to increased conversion rates in this clothing buy-back process. Accordingly, companies might leverage this approach to deduce what works in their context, with their customer base.

Also important to note, our findings are particular to this retailing cash offer scenario. Whether these results will translate to a purchasing, frequently asked questions or healthcare implementation of a chatbot, requires more research. Where anthropomorphism could keep users more engaged in some scenarios, it could also lead to further user frustrations. For example, in a medical diagnosis context, incorporating these anthropomorphic features could inadvertently trigger patients to try and portray themselves in a more positive light (Sproull et al.

1996), and give less accurate depictions of their symptoms. Although anthropomorphism is one aspect that AI designers can use to impact user experience, we also believe that there is fruitful future work evaluating many other aspects like chatbot personality and user based customization.

In summary, our work provides a unique first step toward understanding social and behavioral factors that are worth considering in firms' deployment of autonomous, AI-enabled systems in customer-facing roles. We show that while overall transaction conversion positively increases with anthropomorphism, anthropomorphizing agents can come with several unintended consequences, like greater offer sensitivity. Given that the deployment of chatbots is already quite common, it behooves researchers to further our understanding of best practices for design and implementation of these systems, and what collateral consequences such design decisions may have on the human-agent interaction. It is our hope that this study will spur a new stream of literature in that direction.

Chapter 4

That Sounds Familiar! Dynamic Voice Clones Elicit Greater Trust

4.1 Introduction

Voice-based autonomous agents, such as Google, Siri, and Alexa, are growing more prevalent in many sectors of the economy. Current industry estimates suggest that 38.5% of the United States utilizes voice-based assistants growing every day. [1] Gartner projects that the current market for Voiced Based Personal Assistants is 3.5 billion dollars, with spending associated with direct consumer use, as well as

[1] https://www.emarketer.com/content/voice-assistant-and-smart-speaker-users-2020

employee efficiency tools. [2] As this technology permeates the fabric of both work and everyday life, organizations are beginning to concentrate on how best to craft their audio-based digital presence to create engaging interfaces.

While many organizations have Voice A.I. as top of mind, governmental bodies also have made a concerted effort to block potential deceptive practices. A recent piece of legislation passed by the California State legislature is the Bolstering Online Transparency (B.O.T.) Act (aka "blade-runner bill") which forces transparency into whether A.I. Agents are indeed autonomous by requiring disclosure at the beginning of an interaction. [3] Fines for undisclosed agents can cost an organization up to $2,500 per interaction in the state of California. [4] Additionally, it may only be a matter of time before the U.S. federal government requires the same amount of transparency as the Federal version, the Bot Disclosure and Accountability Act of 2018, recently landed on the Senate floor [5] .

Although the regulatory requirements of transparency help protect the consumer, the efficacy of the act of disclosing A.I. may drastically impact consumer behavior. For instance, in specific contexts like movie and book recommendations (Swearingen and Sinha 2001), or medical diagnosis (Promberger and Baron

[2] https://www.gartner.com/en/newsroom/press-releases/2019-01-09-gartner-predicts-25-percent-of-digital-workers-will-u

[3] shorturl.at/ehDQ4

[4] https://www.dwt.com/blogs/artificial-intelligence-law-advisor/2019/07/is-there-anybody-behind-that-bot

[5] https://www.congress.gov/bill/115th-congress/senate-bill/3127

2006), individuals distrust algorithms and prefer human judgment. However, in other contexts like numeric forecasts, dating recommendations (Logg et al. 2019), or in objective tasks like stock advice (Castelo et al. 2019), individuals prefer algorithms over human judgment. Depending on the organizational context, there may be motives to disclose or not disclose the true nature of the A.I. system.

As much of the literature on algorithmic aversion and appreciation focuses on various contexts, the future use cases of conversational A.I. agents in applications of e-commerce are perhaps more related to economic transaction scenarios. For example, Luo et al. (2019) conducted a study of voice-based chatbots working on behalf of a financial institution, soliciting loan renewals from existing customers. Those authors showed that disclosure of a chatbot's autonomous nature could lead to significant reductions (e.g.79.7 %) in conversion, depending on whether the disclosure is made up-front or after a delay (Luo et al. 2019). Chapter 3 , reported that exposure of an agent's autonomous nature in a text-based sales setting drove significant increases in agent likeability. As these perceptions differ significantly, it remains relatively unclear by what mechanism disclosure influences consumer perceptions. In this work, we consider the implications for trust, as trust is a fundamental aspect of economic exchange (Kosfeld et al. 2005). Disclosure is likely to influence trust because the decision to disclose may foster transparency on the part of a user. Further, failure to explicitly disclose may have negative consequences if a user independently perceives the autonomous nature of the other party.

With the changing legal climate moving in the direction of forced disclosure, organizations may require ways to mitigate the negative appraisals of a fully disclosed agent. One potentially fruitful avenue is personalization. More specifically, a long stream of work speaks to the relationship between homophily and trust, including as it relates to voice-based similarity. While voices have several features(Hildebrand et al. 2020), research in psychology has found that individuals are more attracted to others when they share similar voice characteristics (Dahlbäck et al. 2007). Moreover, other work has shown that voice characteristics can trigger personality inferences. Thus computer-generated voices are preferred when their features imply personality traits that are similar to the user (Nass and Lee 2001). These prior findings are notable because recent technological innovations enable the possibility of voice cloning, in near real-time, on the basis of relatively small voice samples (Jia et al. 2018). With this new technology, there now exists the possibility that organizations can dynamically personalize a voice-based autonomous agent to a particular user in near real-time.

Furthermore, it is essential to consider how disclosure may interact with dynamic personalization. More simply, in the presence of disclosure, we might anticipate that personalization will be perceived positively. For instance, agents that use disclosure may be approached with less scrutiny or cynicism by the user; conversely, absent disclosure, personalization attempts may trigger an adverse reaction if perceived as duplicitous or manipulative. Thus, we explore the following formal research questions in this work: ***To what extent does voice cloning induce trust in an autonomous (spoken) agent? Does disclosure of a spoken agent's autonomous nature causally impact user trust? How does disclosure affect the response to voice-based personalization?***

To address these questions, we build on a version of the behavioral economics "trust game" implemented by Charness and Dufwenberg (2006, 2010), whose studies examined the role of communication on trust. In our adaptation, two players, a human subject and an A.I. agent, play a one-shot game with each other. The A.I. Agent's autonomous nature is randomly disclosed or undisclosed to the human player. After pairing with the agent, the human subject faces a dilemma of whether to trust the other party. Suppose the human player decides to trust the agent. In that case, the automated player can act selfishly, greatly maximizing their payoff or benevolently, significantly increasing the human subject's payoff at a slight expense to their own. Before the human subject makes their decision, the automated agent can randomly send a message. This message is communicated in either a dynamically cloned voice or a default male voice, allowing us to understand how the agent's voice influences the subject's behavior.

Notably, in our game setting, we find that people prefer a cloned version of an A.I. voice compared to a default male voice and no message control. Interestingly, disclosure on its own does not significantly impact end-user trust. However, when examining the interaction of message medium and agent disclosure, we find that dynamic voice cloning, when paired with disclosure, achieves the highest user trust levels. This finding may be driven by subjects noticing that the voices used are indeed automated. When the agent is undisclosed, it represents an act of deception that permanently harms the appraisals of trustworthiness (Schweitzer et al. 2006).

These findings are significant as they contribute to both research and practice. From a theoretical standpoint, this work contributes to the burgeoning work assessing human acceptance of algorithms in human facing job roles (Luo et al. 2019, 2021, Adam et al. 2018, Tambe et al. 2019) by critically examining how individuals trust algorithms in an economic scenario and how aesthetics of these systems can influence behavior. Secondly, we contribute to the literature on personalization in information systems (Adomavicius et al. 2018) and marketing (Ansari et al. 2000, Hosanagar et al. 2014), as we seek to evaluate how a new personalization technique, dynamic voice cloning, could potentially help design effective audio-based experiences. Finally, our research aims to inform practice as we investigate how A.I. disclosure, a regulatory requirement on the horizon, influences human interaction with audio-based conversational agents.

4.2 Literature Review

4.2.1 Algorithms in Human Facing Job Roles

The use of algorithms to support decision-making has been a long-studied line of inquiry (Brunswik 1955, Dawes et al. 1989, Karelaia and Hogarth 2008). As algorithms have become the bedrock of our present society, it is no surprise that human interaction with algorithms is an emerging topic in Marketing (Castelo et al. 2019, Luo et al. 2019), Psychology (Dietvorst et al. 2018, Logg et al. 2019), Computer Science (Yeomans et al. 2019), and Economics (Kleinberg et al. 2017).

One specific stream of research, in this broad area, focuses on human appreciation and aversion to algorithmic judgment. While this stream of work is developing, much of the findings categorize human perceptions of algorithms as highly contextually dependent. In some cases, individuals appreciate algorithms, specifically when dealing with highly objective outcomes like that of numeric forecasts (Logg et al. 2019), providing directions, or predicting the weather (Castelo et al. 2019). In other cases, individuals tend to be more appreciative of human judgment the more subjective the outcome, like recommending music (Yeomans et al. 2019), Human Resource decisions (Castelo et al. 2019, Tambe et al. 2019), and medical diagnoses (Promberger and Baron 2006). As human-facing job roles have a degree of subjectivity to them, it would be no surprise that individuals may have adverse reactions to algorithms in these contexts.

Aligned with the notion that humans react negatively towards A.I. in these roles, (Luo et al. 2019) find that customers dislike interacting with A.I. agents in a sales role. More specifically, they executed a field experiment with a financial loan company in China. A voice-based A.I. agent placed a sales call and asked customers whether they would like to sign up for a loan renewal (Luo et al. 2019). In the outbound calls, the authors randomly manipulate whether to disclose that the agent is autonomous (Luo et al. 2019). This simple act of disclosure reduced loan renewal by 79.7% (Luo et al. 2019). These authors' findings reinforce that there could be drastic real-world consequences to organizations if disclosure is a regulatory requirement.

In contrast, other authors have found that humans can be receptive to A.I. in these roles. In a lab experiment in Chapter 3, we manipulated disclosure of a highly anthropomorphized text-based A.I. Agent in a customer service setting. There we find that disclosure of an A.I. agent's autonomous nature significantly improves its likeability. Additionally, Luo et al. (2021), evaluate how workers in a sales setting react to human versus A.I. sales coaches. In a series of field experiments, sales agents work with potential customers to sell loans and are randomly assigned to receive a human versus A.I. sales coach (Luo et al. 2021). Notably, those authors observed sales improvement for the average salesperson receiving the A.I. coach instead of the human (Luo et al. 2021).

As the emerging literature in human interactions with algorithms in human-facing job roles is fractured and dependent on a myriad of contextual factors, we need a more nuanced approach to better understand these interactions. Behavioral Economics, a methodology that uses games to better understand human decision biases and strategic behavior (e.g. trust, cooperation, negotiation), could help address open questions in this area (Camerer 2019, March 2019, Thaler 2018). While behavioral economics has not focused on studying human-machine interaction, the field has inadvertently studied this phenomenon through the use of computer players in experimental games (March 2019, McCabe et al. 2001, Sanfey et al. 2003). In addition, some Human Computer Interaction (HCI) (e.g., (Kiesler et al. 1996, Torta et al. 2013, Lee et al. 2011)) and Organizational Behavior (Adam et al. 2018) researchers have utilized this methodology to understand human decision making when interacting with A.I. Agents directly.

When individuals interact with computer players in these Behavioral Economics tasks, a common finding is that they behave more rationally and exploitive (Adam et al. 2018, March 2019, Sanfey et al. 2003). More specifically, individuals negotiate less emotionally (Adam et al. 2018, Moretti and Pellegrino 2010, Sanfey et al. 2003), more readily cooperate with their partners (Andreoni and Miller 1993), bid in auctions more conservatively (Teubner et al. 2015), and exploit learned behavior in repeated play (Duersch et al. 2010). The reasoning for the altered behavior is that the need for social consideration when interacting with computer players in these games is not required (March 2019).

While most of these interactions exist through a simple computer interface, humans have proven to behave less predictably when faced with a computer player possessing human-like qualities. For instance, in a prisoner's dilemma game, human subjects tend to reject cooperation with a computer player imbued with a human-like voice compared to a text-based computer or human (Kiesler et al. 1996). Additionally, individuals tend to reason through offers similarly given by a human-like robot as offers made by humans (Torta et al. 2013). These findings are particularly applicable to settings of A.I. agents in human-facing job roles, as their implementation likely utilizes some form of human-like qualities like voice interaction (Luo et al. 2019) or communicative dialog and social cues like in Chapter 3. As humans tend to anthropomorphize A.I. Agents (Moon and Nass 1996), one way to design persuasive interfaces is to look to the social variables identified in human-to-human interactions (Lee et al. 2011).

4.2.2 Similarity Attraction:

Similarity attraction is a phenomenon documented by social psychologists for decades (Byrne 1961). This type of attraction occurs when individuals appraise others as being similar to themselves in some personal dimension like facial attributes (Bailenson et al. 2008), demographics (Banikiotes and Neimeyer 1981, Hu et al. 2008), and attitudes (Singh et al. 2017, Yeong Tan and Singh 1995). Although initially observed in face-to-face interpersonal contexts, similarity-based attraction transfers to several online contexts: social media (Adamopoulos et al. 2018, Aral et al. 2009), computer-mediated communication (Kaptein et al. 2014), and voice-based exchanges (Dahlbäck et al. 2007) to name a few.

HCI designers have also noticed human affinity towards A.I. agents that are similar in some respects to the focal user. More specifically, individuals trust A.I. agents that possess similar paralinguistic vocal cues that align to their personality (Nass and Lee 2001) and display accents that are congruent with their own (Dahlbäck et al. 2007). In a recent experiment investigating human A.I. teamwork, Trainer et al. (2020) found that people prefer to work with A.I. agents whose avatar image displayed similar racial and gender characteristics and ultimately trusted these agents more than those that were dissimilar. As they relate to A.I. Agent design, these findings show that individuals tend to trust systems that display similar aesthetics to themselves.

Although there are many reasons for similarity attraction, a dominant explanation for the phenomenon is the reinforcement-affect model (Byrne and Clore 1970). This model postulates that people enjoy experiencing positive stimuli (Byrne and Clore 1970) and actively seek them. Since interactions with similar individuals tend to go more smoothly, as opposed to dissimilar stimuli, individuals are drawn toward stimuli congruent with these attributions (Berscheid and Hatfield 1978). For example, when individuals hear a similar voice, they associate the sound with subsequent positive conversations, as there is already a basis for common ground. That said, it is somewhat unclear how disclosure of an A.I. agent will interact with similarity-based aesthetics. On the one hand, disclosure alone may be enough to push individuals away from the A.I. agent as it highlights dissimilarity. However, on the other, it may mitigate the initial negative dissimilarity appraisals.

As there are reasons to believe that A.I. agent disclosure will influence trust, and similar voice aesthetics may mitigate negative aspects invoked by disclosure, we look to a behavioral economics experiment to help us evaluate these features.

4.3 Methods:

4.3.1 Experiment Design

To evaluate our research questions, we utilize a Behavioral Economics game introduced by Charness and Dufwenberg (2006), which specifically addresses how communication influences trust. We chose this specific game, as it is, to our knowledge, the most widely used game to evaluate communication's impact on trust, as other trust games prohibit the use of pre-play communication (Berg et al. 1995). In the original game design, articulated in Figure 4.1, subjects are initially "matched" to a playing partner and randomly assigned to play the role of agent A or B. In a manipulation condition, subjects playing the role of B would have the opportunity to pass a written message to party A, stating whatever they like, e.g., superfluous comments, promises, etc. Subsequently, party A would decide whether they wished to play the game with B (choosing "in" or "out"). If party A decided to play, B would choose to walk away with some money, leaving A with nothing, or roll a dice. Choosing to roll, the final payoff to both parties would depend on the resulting number that was rolled. This study was originally designed to examine how message passing would influence the decisions of both A (whether to trust B), and B (how to behave upon receiving that trust).

We modified this game, implementing it in a digital setting, and had the role

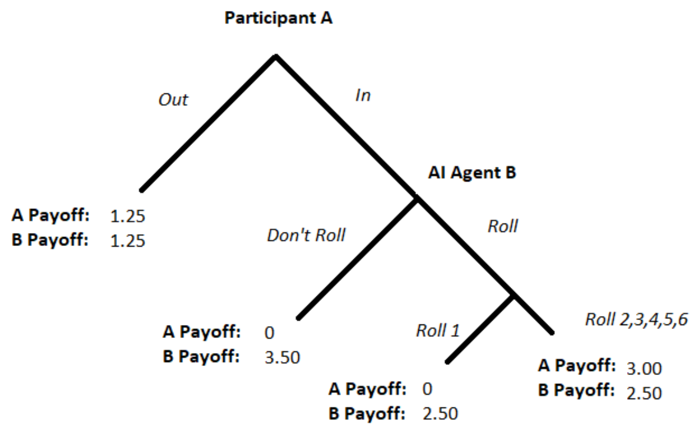

Figure 4.1: Payoffs Adapted from Charness & Dufwenberg 2006

of party B always played by an autonomous agent. Under this setup, we had two conditions that mirrored the original design of the study. In a control condition, no communication would be passed from party B to party A. In a (written) message condition, party A would receive a written text-based message from the agent, where the message is drawn from those exchanged in the original studies of Charness and Dufwenberg (2006, 2010). Beyond these conditions, we incorporated a series of other experimental conditions. We added conditions wherein the communication would take place via a voice-based message, i.e., a recording. Further, we manipulated the voice used in generating that recording, employing a 'default' computer voice in one condition and a dynamic voice clone (based on the recorded consent statement) in another condition. Finally, we implemented a series of conditions that mirrored Figure 4.1, except that we also disclosed the autonomous nature of party B.

CD 2006 Control (No Message)	CD 2006 Text (Written)	Control + Disclosure	Default Voice (Verbal Message)	Default Voice + Disclosure	Cloned Voice (Verbal Message)	Cloned Voice + Disclosure

Table 4.1: Experiment Conditions

A detailed outline of how subjects play this behavioral economics game can be found in Figure 4.2 , and depicts the three-step process. First, a subject calls into a 1-800 number and reads a consent statement over the phone. Next, a server records the consent statement, randomizes the subject into one of the experimental conditions, and returns a unique completion code. Third, the subject then enters the completion code and plays the trust game through a computer.

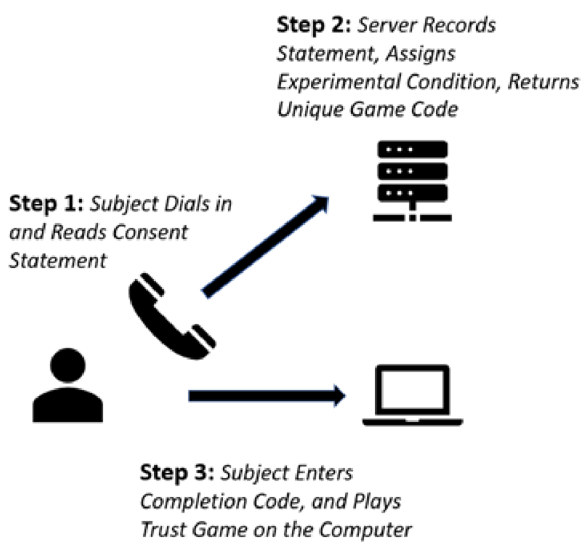

Figure 4.2: Study Enrollment Process

4.3.2 Voice Cloning:

To operationalize voice cloning in the experimental design, we utilized an open-source implementation of Jia et al. (2018). This voice cloning method uses transfer learning, a Deep Learning technique, which allows for learning in one task to transfer over to the learning in another (Pan and Yang 2010) . The voice cloning model is three parts: the Speaker Encoder, Synthesizer, and Vocoder. The Speaker Encoder takes the brief audio file and generates a speaker embedding. The embedding is used with the Synthesizer, which maps the to-be-generated spoken words to a mel-spectrogram. This spectrogram is fed into the Vocoder, which takes the spectrogram image and generates an audio file. An illustration of Jia et al. (2018), 2018's three-part model can be found in Figure 4.3 .

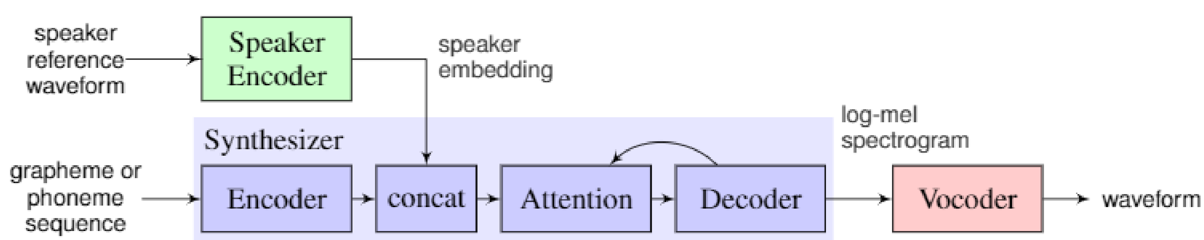

Figure 4.3: Jia et al 2018 Voice Cloning Model

The voice cloning software was installed on a GPU instance on Amazon Web Services. When a subject, randomized into a voice clone condition, reads their consent statement, this pipeline is triggered. The resulting audio is loaded onto an S3 bucket, which would be available as a message from Player B during gameplay.

4.4 Empirical Approach & Data

To address our research questions of how both disclosure and voice cloning influence user trust in A.I. agents, and to what extent these two features interact, we look to a controlled online lab experiment conducted on Prolific. Our outcome variable of interest is whether the subject chose to be "in" and trust player B during gameplay; this denotes the binary indicator variable in. Additionally, to tease out the effects of disclosure, voice cloning, and their interaction with each other, we utilize treatment dummies, Disclosure, Message, and Condition, respectively. Disclosure indicates whether Player B is displayed to the study participant as "Participant B" or "Automated Agent B." Message is how the message is sent to the human subject (e.g. voice clone, default voice, or no message). Condition relates to the overall treatment dummy associated with the treatment condition in Table 4.1. For this analysis, we execute a series of logistic regressions to identify how these treatments influence trust, regressing the binary outcome variable in on the various treatment dummies.

4.4.1 Power Analysis:

Before conducting the experiment, a power analysis was executed, replicating Charness and Dufwenberg (2006), with a .8 probability of detecting an effect size associated with the message versus no message (e.g., 18%) at an alpha of .05 we would need to recruit at least 109 subjects per condition.

4.4.2 Subject Recruitment:

For this experiment, we recruited 1,118 subjects through Prolific. We exceeded the minimum threshold of subjects of 109 for the control conditions (i.e., CD Control, and Control with Disclosure) to limit the load on our servers as generating too many audio messages in a short period would cause technical issues. Additionally, we limited the subject pool to individuals from the United States and Canada as the A.I. models used to produce the recorded messages only generated audio in North American English accents. Limiting the subject pool in this way prevented additional accent based similarity attributions which could bias the results (Dahlbäck et al. 2007).

4.4.3 Descriptive Statistics

Overall, shown in Table 4.2, participants trusted player B on average 75.4% of the time. They also spent 322 seconds (about 5 mins) playing the game. Subjects, on average, received .410 of the disclosure treatment, as the CD condition did not include disclosure. For randomization into receiving a message, this was close to an even split.

Table 4.2: Descriptive Statistics

Variable	**Mean**	**St. Dev.**	**Min**	**Max**
in rate	0.754	0.431	0.00	1.00
Total Time(sec)	322	194	48	263
Disclosure	0.410	0.492	0.00	1.00
Message	.493	0.50	0.00	1.00
Age	33.4	11.8	18	78

Breakdowns showing the number of subjects per condition are in Table 4.3. To reiterate, more subjects are in the control conditions as generating audio messages was highly taxing on the server, so for each treatment, we added a control No Message condition.

Table 4.3: Subjects per Condition

CD 2006 Control (No Message)	**CD 2006 Text (Written)**	**Control + Disclosure**	**Default Voice (Verbal Message)**	**Default Voice + Disclosure**	**Cloned Voice (Verbal Message)**	**Cloned Voice + Disclosure**
N=331	**N=109**	**N=236**	**N=108**	**N=107**	**N=112**	**N=115**

4.5 Results

4.5.1 Message Type:

To address our first research question, how does voice cloning influence trust, we look to compare the conditions by mode of messages sent by player B. As shown in Figure 4.4, which depicts in rate by message type. Here we see that the Clone condition has the highest in rate at 86.34%. To understand if these differences are significantly different, we run a logistic regression, output shown in Table 4.4 with the Clone voice conditions as the base condition to compare against the other three message types.

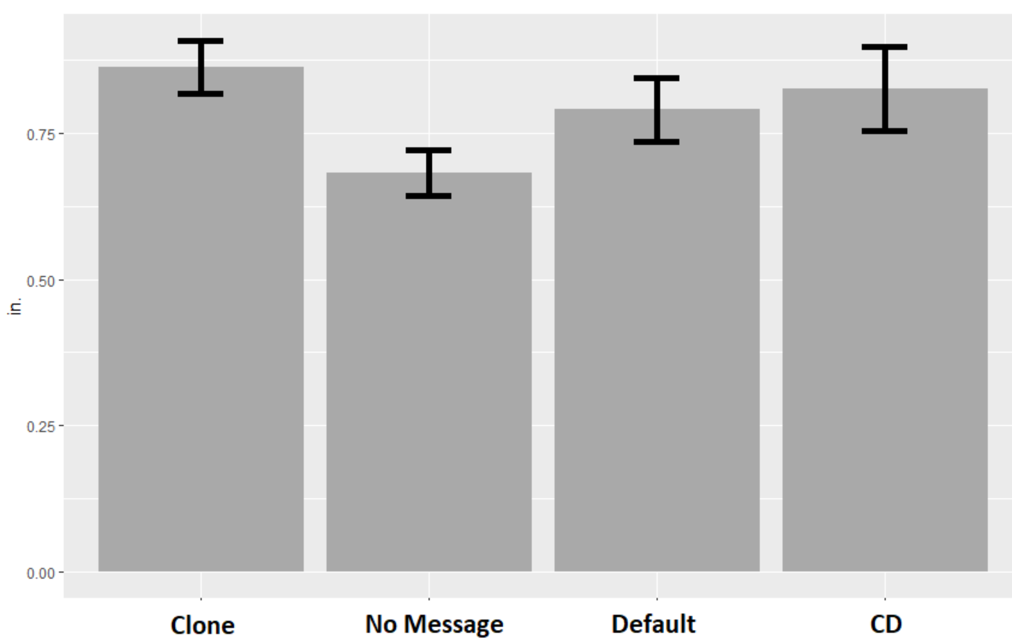

Figure 4.4: In Rate by Message Type - (95% Confidence Intervals)

In comparison to the No Message and Default Voice conditions, voice cloning appears to perform significantly better in eliciting user trust at (p≤ 001) and (p≤ 01). For the comparison (Charness and Dufwenberg 2006), voice cloning does not perform significantly better than the text-based benchmark. These results show that voice cloning, regardless of disclosure, appears to be the best mode of interaction when spoken messages are exchanged.

Table 4.4: Logistic Regression Comparing Message Type (Clone as Reference)

DV = In	**Coefficient**
No Message	-1.0787*** (0.2133)
Default Voice	-0.5150** (0.2559)
CD Text	-0.2888 (0.3180)
Observations	1,118

*** $p \leq .001$, ** $p \leq 0.01$

4.5.2 Disclosure vs No Disclosure

Next, we look to understand how overall disclosure impacts end-user trust. In looking at Figure 4.5, we see that the No Disclosure and Disclosure conditions are nearly identical with in rates of 75.00% and 75.9%, respectively. After running a logistic regression to compare the two groups, as expected, the differences are not statistically significant.

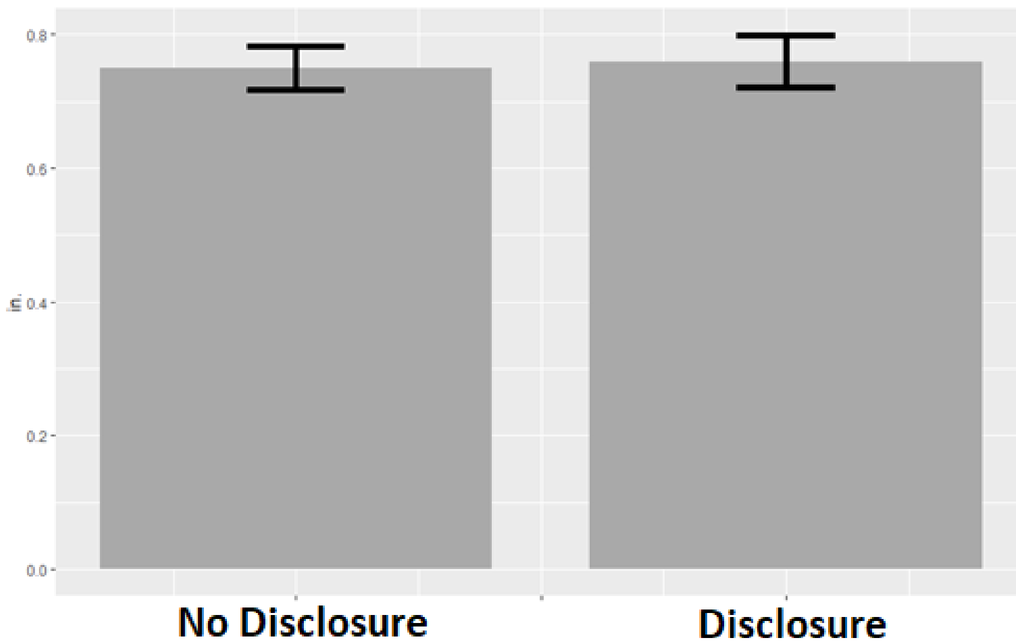

Figure 4.5: In Rate by Disclosure - (95% Confidence Intervals)

4.5.3 Overall Results

Finally, to better understand how both disclosure and message type interplay, we compare all seven conditions against each other. In Figure 4.6 depicts the in rate for each condition. Interestingly, the clone with disclosure condition achieves the highest elicitation of human trust with an in rate of 89.5%. Notably, the disclosure with clone condition is driving most of the in rate for the message analysis above. To identify how in rates for each of these experimental conditions compare to each other, we permute a series of logistic regressions in Table 4.5.3, changing the base condition. In Model 7, the clone with disclosure condition appears to outperform both the control from Charness and Dufwenberg (2006), and control with disclosure condition at ($p \leq 01$) significance. Next, when comparing the clone with disclosure to the default voice, the clone with disclosure outperforms both versions of the default voice at ($p \leq 05$) significance. Finally, for the clone without disclosure and CD conditions, the clone with disclosure does perform directionally better than both conditions; however, the difference is not significant at the($p \leq 05$) level, with ($p \leq 13$) and ($p \leq 15$), respectively.

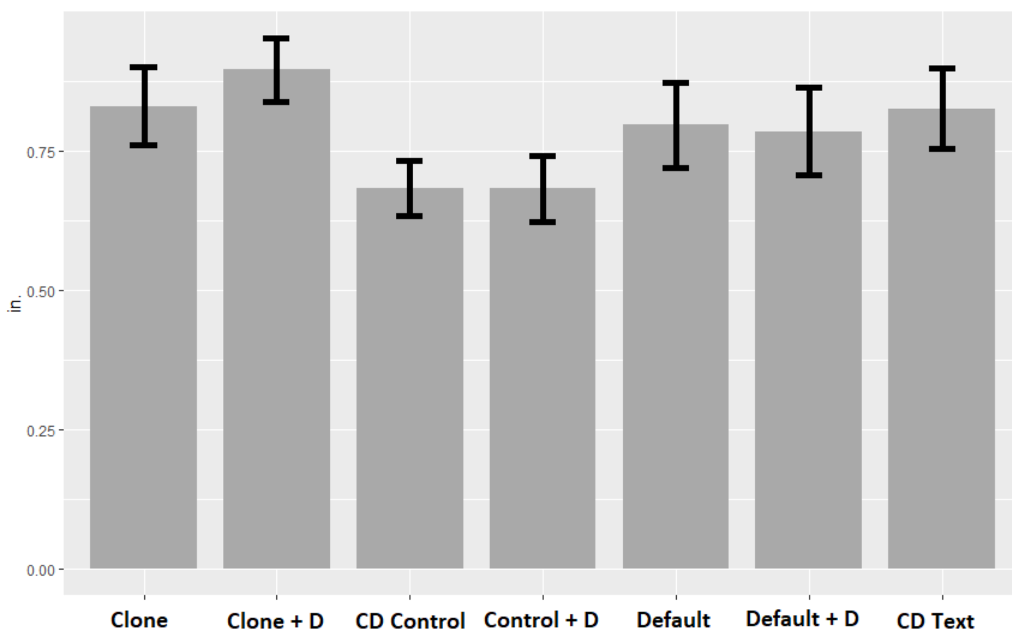

Figure 4.6: In Rate by Condition - (95% Confidence Intervals)

Table 4.5: Logistic Regression All Conditions

	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7
(Intercept)	0.77 ***	0.76 ***	1.36 ***	1.30 ***	1.56 ***	1.59 ***	2.15 ***
	(0.12)	(0.14)	(0.24)	(0.24)	(0.25)	(0.25)	(0.31)
Control + D	-0.00		-0.60 **	-0.53 *	-0.79 ***	-0.82 ***	-1.39 ***
	(0.18)		(0.28)	(0.27)	(0.29)	(0.29)	(0.34)
Default Voice	0.60 **	0.60 **		0.07	-0.19	-0.22	-0.79 **
	(0.27)	(0.28)		(0.34)	(0.35)	(0.35)	(0.39)
Default + D	0.53 **	0.53 *	-0.07		-0.26	-0.29	-0.85 **
	(0.26)	(0.27)	(0.34)		(0.35)	(0.34)	(0.39)
CD Text	0.79 ***	0.79 ***	0.19	0.26		-0.03	-0.59
	(0.28)	(0.29)	(0.35)	(0.35)		(0.36)	(0.40)
Clone	0.82 ***	0.82 ***	0.22	0.29	0.03		-0.56
	(0.28)	(0.29)	(0.35)	(0.34)	(0.36)		(0.40)
Clone + D	1.38 ***	1.39 ***	0.79 **	0.85 **	0.59	0.56	
	(0.33)	(0.34)	(0.39)	(0.39)	(0.40)	(0.40)	
CD Control		0.00	-0.60 **	-0.53 **	-0.79 ***	-0.82 ***	-1.38 ***
		(0.18)	(0.27)	(0.26)	(0.28)	(0.28)	(0.33)
AIC	1223.04						
BIC	1258.17						
Log Likelihood	-604.52						
Deviance	1209.04						
Num. obs.	1118						

*** $p \leq .01$, ** $p \leq 0.05$

4.5.4 Randomization Checks

In our experimental setup, we randomized individuals when they entered the experiment. To ensure that the randomization process occurred successfully, we execute a series of pairwise t-tests to determine if observable characteristics are balanced. We do this for both age and location latitude. As shown in Table 4.6 and Table 4.7, there are no significant differences between groups in geography and age. From these results, we can conclude that randomization was executed successfully.

Table 4.6: Pairwise t-Tests evaluating Latitude Balance per Condition

	Clone	**Clone + D**	**CD Control**	**Control + D**	**Default**	**Default + D**
Clone + D	1.00					
CD Control	1.00	1.00				
Control + D	1.00	0.76	1.00			
Random	1.00	1.00	1.00	0.88		
Random + D	1.00	1.00	1.00	1.00	1.00	
CD Text	1.00	1.00	1.00	1.00	1.00	1.00

Table 4.7: Pairwise t-Tests evaluating Age Balance per Condition

	Clone	Clone + D	CD Control	Control + D	Default	Default + D
Clone + D	1.00					
CD Control	1.00	1.00				
Control + D	1.00	1.00	1.00			
Random	1.00	1.00	1.00	1.00		
Random + D	1.00	1.00	1.00	1.00	1.00	
CD Text	1.00	1.00	1.00	1.00	1.00	1.00

4.6 Discussion

Our work provides a novel first look at a potentially impactful design aesthetic of spoken A.I. agents, dynamic voice cloning. Our behavioral economics experiment shows that incorporating voice cloning may be a viable option to elicit trust from individuals interacting with A.I. agents in collaborative and economic interactions.

Secondly, this work evaluates how A.I. Agent disclosure influences trust in A.I. agents. While previous fieldwork suggests that individuals react adversely when agents disclose that they are automated (Luo et al. 2019), from our study, we find that disclosure by itself does not necessarily negatively impact trust. That said, some of the discrepancies between Luo et al. (2019), could be that the "realism" of voices utilized in our experiment is lower than theirs. We used recordings from lower quality telephone calls, thus making the generated output somewhat rough and less lifelike. In turn, users may readily identify that the agent is not a human and breaking any initial trust perceptions a subject may have. Regardless of the voice, mending this broken trust may prove difficult (Schweitzer et al. 2006). It may also be a cautionary tale for organizations that look to pursue a non-disclosure strategy, as it may destroy consumer trust at the onset.

Perhaps a more realistic comparison to a true human actor, in our experiment, would be the text-based CD condition, as text messaging has few social cues for subjects to determine if the agent is indeed human (Walther 1992, Walther and Tidwell 1995). That said, our dynamic voice clone with disclosure condition performs at least as well as the CD condition. This result furthers the notion that a potentially more viable strategy for organizations to pursue is voice-based personalization with disclosure, as there is no benefit to withholding this information.

Thirdly, we evaluated how disclosure interacts with dynamic voice cloning. From our results, we see that disclosure, when paired with voice cloning, achieves the highest in rate to a statistically significant degree over the default voice (with or without disclosure) and no message conditions. These findings are interesting as it shows a potential path forward for organizations if faced with increased regulatory disclosure requirements. Additionally, further data requirements and regulations, outside of A.I. disclosure, like GDPR, could be on the horizon , limiting what archival data organizations can use to personalize their customers' experience(Sun et al. 2020). Dynamic voice cloning can be done on the fly with a small clip of a customer's voice without needing to persist any user data.

While our research evaluates a potential new way for organizations to personalize their audio-based experiences, our work has several limitations. Firstly, our experimental context occurs in a controlled online lab experiment; it is unclear how much trust is involved in outcome variables of interest for marketers and organizations (e.g., sales conversions, lead generation). Although we feel that the lab-based setting gives our research the ability to tease out the potential influence of disclosure on trust, further work should be done in the field evaluating dynamic voice cloning's effectiveness in the real world.

Another limitation of our work is its focus on North American English speakers. It is not clear if our results necessarily generalize across cultures. More notably, where Luo et al. (2019) takes place in a field setting in China, which finds that disclosure negatively impacts loan renewals. Further work on A.I. Agent disclosure could seek to evaluate if there are perhaps country-based differences in the reaction to this disclosure and voice-based aesthetic.

4.7 Conclusion

In conclusion, this work seeks to help A.I. practitioners evaluate two crucial features, A.I. disclosure, and voice-based aesthetic. While one feature, A.I. Disclosure, may be a regulatory requirement for all A.I. Agents in the future, we find that dynamic voice cloning may help to elicit higher levels of trust in disclosed voice-based systems. As voice-based A.I. is becoming an ever more critical consumer channel, dynamic voice cloning could be a potentially fruitful step forward in designing transparent voice-based consumer experiences.

Chapter 5

Closing Remarks

In this work, I present three chapters investigating how humanization tactics, when imbued into interactions between digital entities and users, can greatly influence human behavior. These humanization tactics have proven effective via experiments in three digital mediums Social Media, Messaging Platforms, and Voice A.I.. Additionally, these chapters show that humanizing strategies effectively influence various outcomes of interest for organizations (e.g., sales conversion, price sensitivity, and trust).

In my first essay, I illustrate how an organization can deploy Politeness strategies in a Social Media advertising context to drive higher conversion rates. This work also highlights that appropriate use of language in a social media context is driven by the type of relationship an organization has with the focal customer.

Notably, this finding mirrors that found in human-to-human request behavior (Brown and Levinson 1987). My second essay furthers this line of inquiry by studying how the intensity of anthropomorphism in a text-based chatbot, deployed in a customer service setting, influences conversion and price sensitivity. Interestingly, we find that while conversion increases with anthropomorphism, this comes at the cost of higher consumer price sensitivity. This finding furthers the notion that making A.I. interactions more human-like is not always the best design choice (Mathur and Reichling 2016). Finally, my third essay seeks to evaluate two important voice-based A.I. design characteristics disclosure and voice personalization and their influence on trust. Disclosure is critical to consider as incoming legislation forces organizations to disclose A.I. bots as automated at the onset of an interaction, a requirement that could adversely impact sales (Luo et al. 2019). As disclosure may be an inevitability, organizations need ways to better design voice-based interactions to overcome potential negative aspects of this requirement. As such, we find that dynamic voice cloning is a potentially viable option to increase levels of trust with these disclosed AI agents.

In the future, organizations will continue to be inundated with messages from their customers. At first blush this problem is one which should be approached with mechanistic efficiency, but upon further review this work highlights the need for humanization of these interactions. To guarantee success in a more conversational future world it is paramount that organizations focus on their human touch.

References Section

References

Aaker, J. L. (1997). Dimensions of brand personality. *SSRN Electronic Journal*.

Adam, M. T., Teubner, T., and Gimpel, H. (2018). No Rage Against the Machine: How Computer Agents Mitigate Human Emotional Processes in Electronic Negotiations. *Group Decision and Negotiation*, 27(4):543–571.

Adamopoulos, P., Ghose, A., and Todri, V. (2018). The impact of user personality traits on word of mouth: Text-mining social media platforms. *Information Systems Research*, 29(3):612–640.

Adomavicius, G., Bockstedt, J. C., Curley, S. P., and Zhang, J. (2018). Effects of online recommendations on consumers' willingness to pay. *Information Systems Research*, 29(1):84–102.

Aggarwal, P. (2004). The effects of brand relationship norms on consumer attitudes and behavior. *Journal of Consumer Research*, 31(1):87–101.

Aggarwal, P. and Mcgill, A. L. (2012). When brands seem human, do humans act like brands? automatic behavioral priming effects of brand anthropomorphism.

Journal of Consumer Research, 39(2):307–323.

Ai, C. and Norton, E. C. (2003). Interaction terms in logit and probit models. *Economics Letters*, 80(1):123–129.

Andreoni, J. and Miller, J. (1993). Rational Cooperation in Finitely Repeated Prisoner's Dilemma: Experimental Evidence. 103:570–585.

Ansari, A., Essegaier, S., and Kohli, R. (2000). Internet recommendation systems. *Journal of Marketing Research*, 37(3):363–375.

Applegate, E. (2016). *Strategic copywriting: how to create effective advertising*. Rowman & Littlefield.

Aral, S., Muchnik, L., and Sundararajan, A. (2009). Distinguishing influence-based contagion from homophily-driven diffusion in dynamic networks. *Proceedings of the National Academy of Sciences of the United States of America*, 106(51):21544–21549.

Araujo, T. (2018). Living up to the chatbot hype: The influence of anthropomorphic design cues and communicative agency framing on conversational agent and company perceptions. *Computers in Human Behavior*, 85:183–189.

Bailenson, J. N., Iyengar, S., Yee, N., and Collins, N. A. (2008). Facial similarity between voters and candidates causes influence. *Public Opinion Quarterly*, 72(5):935–961.

Banikiotes, P. G. and Neimeyer, G. J. (1981). Construct importance and rating similarity as determinants of interpersonal attraction. *British Journal of Social Psychology*, 20:259–263.

Bapna, S., Benner, M. J., and Qiu, L. (2017). Nurturing online communities: An empirical investigation. *SSRN Electronic Journal*.

Barrett, J. L. and Keil, F. C. (1996). Conceptualizing a nonnatural entity: Anthropomorphism in god concepts. *Cognitive Psychology*, 31(3):219–247.

Bartneck, C., Kulić, D., Croft, E., and Zoghbi, S. (2008). Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots. *International Journal of Social Robotics*, 1(1):71–81.

Bayan, R. (2006). *Words that sell: more than 6,000 entries to help you promote your products, services, and ideas*. McGraw-Hill.

Berg, J., Dickhaut, J., and McCabe, K. (1995). Trust, reciprocity, and social history.

Berscheid, E. and Hatfield, E. (1978). *Interpersonal attraction*. Reading, Mass. : Addison-Wesley Pub. Co., Reading, Mass., 2d ed.. edition.

Blum-Kulka, S. (1987). Indirectness and politeness in requests: Same or different? *Journal of Pragmatics*, 11(2):131–146.

Brehm, J. W. (1966). *A Theory of Psychological Reactance*. Academic Press.

Brown, P. and Levinson, S. C. (1987). *Politeness: some universals in language use*. Cambridge Univ. Pr.

Brunswik, E. (1955). Representative design and probabilistic theory in a functional psychology. *Psychological Review*, 62(3):193–217.

Byrne, D. (1961). Interpersonal attraction and attitude similarity. *The Journal of Abnormal and Social Psychology*, 62(3):713–715.

Byrne, D. and Clore, G. L. (1970). A reinforcement model of evaluative responses.

Camerer, C. F. (2019). Artifi cial Intelligence and Behavioral Economics. (May):587–608.

Caporael, L. (1986). Anthropomorphism and mechanomorphism: Two faces of the human machine. *Computers in Human Behavior*, 2(3):215–234.

Cassell, J. and Bickmore, T. (2000). External manifestations of trustworthiness in the interface. *Communications of the ACM*, 43(12):50–56.

Castelo, N., Bos, M. W., and Lehmann, D. R. (2019). Task-Dependent Algorithm Aversion. *Journal of Marketing Research*, 56(5):809–825.

Charness, G. and Dufwenberg, M. (2006). Promises and partnership. *Econometrica*, 74(6):1579–1601.

Charness, G. and Dufwenberg, M. (2010). Bare promises: An experiment. *Economics Letters*, 107(2):281–283.

Chen, K.-J., Lin, J.-S., Choi, J. H., and Hahm, J. M. (2015). Would you be my friend? an examination of global marketers brand personification strategies in social media. *Journal of Interactive Advertising*, 15(2):97–110.

Clark, M. S. and Mils, J. (1993). The difference between communal and exchange relationships: What it is and is not. *Personality and Social Psychology Bulletin*, 19(6):684–691.

Crozier (2017). Lufthansa delays chatbot's responses to make it more 'human'.

Daft, R. L. and Lengel, R. H. (1986). Organizational information requirements, media richness and structural design. *Management science*, 32(5):554–571.

Dahlbäck, N., Wang, Q., Nass, C., and Alwin, J. (2007). Similarity is more important than expertise: Accent effects in speech interfaces. *Conference on Human Factors in Computing Systems - Proceedings*, pages 1553–1556.

Dale, R. (2016). The return of the chatbots. *Natural Language Engineering*, 22(05):811–817.

Danaher, P. J., Smith, M. S., Ranasinghe, K., and Danaher, T. S. (2015). Where, when, and how long: Factors that influence the redemption of mobile phone coupons. *Journal of Marketing Research*, 52(5):710–725.

Dawes, R. M., Faust, D., and Meehl, P. E. (1989). Clinical Versus Actuarial Judgement. *Science*, 243(4899):1668.

Deci, E. and Ryan, R. M. (1985). *Intrinsic motivation and self-determination in human behavior*. Springer Science & Business Media.

Deke, J. et al. (2014). Using the linear probability model to estimate impacts on binary outcomes in randomized controlled trials. Technical report, Department of Health and Human Services.

Dietvorst, B. J., Simmons, J. P., and Massey, C. (2015). Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General*, 144(1):114–126.

Dietvorst, B. J., Simmons, J. P., and Massey, C. (2018). Overcoming algorithm aversion: People will use imperfect algorithms if they can (even slightly) modify them. *Management Science*, 64(3):1155–1170.

Dillard, J. P. (2007). *The persuasion handbook: developments in theory and practice.* Sage.

Dillard, J. P. and Shen, L. (2005). On the nature of reactance and its role in persuasive health communication. *Communication Monographs*, 72(2):144–168.

Dolen, W. M. V., Ruyter, K. D., and Streukens, S. (2008). The effect of humor in electronic service encounters. *Journal of Economic Psychology*, 29(2):160–179.

Don, A., Brennan, S., Laurel, B., and Shneiderman, B. (1992). Anthropomorphism: from eliza to terminator 2. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 67–70. ACM.

Duersch, P., Kolb, A., Oechssler, J., and Schipper, B. C. (2010). Rage against the machines: How subjects play against learning algorithms. *Economic Theory*, 43(3):407–430.

Duffy, B. R. (2003). Anthropomorphism and the social robot. *Robotics and Autonomous*

Systems, 42(3-4):177–190.

Eckles, D., Gordon, B. R., and Johnson, G. A. (2018). Field studies of psychologically targeted ads face threats to internal validity. *Proceedings of the National Academy of Sciences*, page Article in Advance.

Firth, D. (1993). Bias reduction of maximum likelihood estimates. *Biometrika*, 80(1):27.

Fitzsimons, G. J. and Lehmann, D. R. (2004). Reactance to recommendations: When unsolicited advice yields contrary responses. *Marketing Science*, 23(1):82–94.

Fournier, S. (1998). Consumers and their brands: Developing relationship theory in consumer research. *Journal of Consumer Research*, 24(4):343–353.

Francis, L., Monahan, K., and Berger, C. (1999). A laughing matter? the uses of humor in medical interactions. *Motivation and Emotion*, 23:155–174.

Frei, E., Elias, A., Wheeler, C., Richardson, P., and Hryniuk, W. (1998). The relationship between high-dose treatment and combination chemotherapy: the concept of summation dose intensity. *Clinical cancer research*, 4(9):2027–2037.

Fussell, S. R., Kiesler, S., Setlock, L. D., and Yew, V. (2008). How people anthropomorphize robots. *Proceedings of the 3rd international conference on Human robot interaction - HRI 08*.

Gefen, D. and Straub, D. W. (2003). Managing user trust in b2c e-services. *e-Service Journal*, 2(2):7–24.

Gnewuch, U., Morana, S., Adam, M., and Maedche, A. (2018). Faster is not always better: Understanding the effect of dynamic response delays in human-chatbot interaction. *Twenty-Sixth European Conference on Information Systems.*

Goetz, J., Kiesler, S., and Powers, A. (2003). Matching robot appearance and behavior to tasks to improve human-robot cooperation. *The 12th IEEE International Workshop on Robot and Human Interactive Communication, 2003. Proceedings. ROMAN 2003.*

Goffman, E. (1967). *On Face-Work. An Analysis of Ritual Elements in Social Interaction.* Doubleday.

Goh, K. Y., Heng, C. S., and Lin, Z. (2012). Social media brand community and consumer behavior: Quantifying the relative impact of user- and marketer-generated content. *SSRN Electronic Journal.*

Goldfarb, A. (2015). *Economic analysis of the digital economy.* Univ. of Chicago Press.

Gordon, B. R., Zettelmeyer, F., Bhargava, N., and Chapsky, D. (2017). A comparison of approaches to advertising measurement: Evidence from big field experiments at facebook. *SSRN Electronic Journal.*

Gurth, W., Schmittberger, R., and Schwarze, B. (1982). An experimental analysis of ultimatum bargaining. *An experimental analysis of ultimatum bargaining*, 3(4):367–388.

Guthrie, S. E. (1995). *Faces in the clouds a new theory of religion.* Oxford University Press.

Heider, F. and Simmel, M. (1944). An experimental study of apparent behavior. *The American Journal of Psychology*, 57(2):243.

Hildebrand, C., Efthymiou, F., Busquet, F., Hampton, W. H., Hoffman, D. L., and Novak, T. P. (2020). Voice analytics in business research: Conceptual foundations, acoustic feature extraction, and applications. *Journal of Business Research*, 121(January):364–374.

Holtgraves, T. (2011). *Language as social action social psychology and language use.* Routledge.

Holtgraves, T. and Han, T.-L. (2007). A procedure for studying online conversational processing using a chat bot. *Behavior Research Methods*, 39(1):156–163.

Holtgraves, T., Ross, S., Weywadt, C., and Han, T. (2007). Perceiving artificial social agents. *Computers in Human Behavior*, 23(5):2163–2174.

Holtgraves, T. and Yang, J.-n. (1990). Politeness as universal: Cross-cultural perceptions of request strategies and inferences based on their use. *Journal of Personality and Social Psychology*, 59(4):719–729.

Holzwarth, M., Janiszewski, C., and Neumann, M. M. (2006). The influence of avatars on online consumer shopping behavior. *Journal of Marketing*, 70(4):19–36.

Horrace, W. C. and Oaxaca, R. L. (2006). Results on the bias and inconsistency of ordinary least squares for the linear probability model. *Economics Letters*, 90(3):321–327.

Hosanagar, K., Fleder, D., Lee, D., and Buja, A. (2014). Will the global village fracture into tribes recommender systems and their effects on consumer fragmentation. *Management Science*, 60(4):805–823.

Hu, C., Thomas, K. M., and Lance, C. E. (2008). Intentions to initiate mentoring relationships: Understanding the impact of race, proactivity, feelings of deprivation, and relationship roles. *Journal of Social Psychology*, 148(6):727–744.

Jain, H., Padmanabhan, B., Pavlou, P. A., and Santanam, R. T. (2018). Call for papers—special issue of information systems research—humans, algorithms, and augmented intelligence: The future of work, organizations, and society. *Information Systems Research*, 29(1):250–251.

Jaworski, A. and Coupland, N. (2014). *The discourse reader*. Routledge.

Jia, Y., Zhang, Y., Weiss, R. J., Wang, Q., Shen, J., Ren, F., Chen, Z., Nguyen, P., Pang, R., Moreno, I. L., and Wu, Y. (2018). Transfer learning from speaker verification to multispeaker text-to-speech synthesis. *Advances in Neural Information Processing Systems*, 2018-December(NeurIPS):4480–4490.

Kaiser, J. and Lacy, M. (2009). A general-purpose method for two-group randomization tests. *The Stata Journal*.

Kaptein, M., Castaneda, D., Fernandez, N., and Nass, C. (2014). Extending the similarity-attraction effect: The effects of when-similarity in computer-mediated communication. *Journal of Computer-Mediated Communication*, 19(3):342–357.

Karelaia, N. and Hogarth, R. M. (2008). Determinants of Linear Judgment: A Meta-Analysis of Lens Model Studies. *Psychological Bulletin*, 134(3):404–426.

Kennedy, J. S. (2003). *The new anthropomorphism*. Lightning Source UK.

Kiesler, S., Powers, A., Fussell, S. R., and Torrey, C. (2008). Anthropomorphic interactions with a robot and robot–like agent. *Social Cognition*, 26(2):169–181.

Kiesler, S., Siegel, J., and McGuire, T. W. (1984). Social psychological aspects of computer-mediated communication. *American psychologist*, 39(10):1123.

Kiesler, S., Sproull, L., and Waters, K. (1996). A prisoners dilemma experiment on cooperation with people and human-like computers. *Journal of Personality and Social Psychology*, 70(1):47–65.

Kleinberg, J., Lakkaraju, H., Leskovec, J., Ludwig, J., and Mullainathan, S. (2017). Human decisions and machine predictions*. *The Quarterly Journal of Economics*.

Kosfeld, M., Heinrichs, M., Zak, P. J., Fischbacher, U., and Fehr, E. (2005). Oxytocin increases trust in humans. *Nature*, 435(7042):673–676.

Kronrod, A., Grinstein, A., and Wathieu, L. (2012a). Enjoy! hedonic consumption and compliance with assertive messages. *Journal of Consumer Research*, 39(1):51–61.

Kronrod, A., Grinstein, A., and Wathieu, L. (2012b). Go green! should environmental messages be so assertive? *Journal of Marketing*, 76(1):95–102.

Kwon, E. S. and Sung, Y. (2011). Follow me! global marketers' twitter use. *Journal of Interactive Advertising*, 12(1):4–16.

Lance, C. E. (1988). Residual centering, exploratory and confirmatory moderator analysis, and decomposition of effects in path models containing interactions. *Applied Psychological Measurement*, 12(2):163–175.

Lee, D., Hosanagar, K., and Nair, H. (2017). Advertising content and consumer engagement on social media: Evidence from facebook. *Management Science*.

Lee, M. K., Kiesler, S., and Forlizzi, J. (2011). 2011-CHI-behavioral_econ. pages 325–334.

Liebman, N. and Gergle, D. (2016). It-s (not) simply a matter of time: The relationship between cmc cues and interpersonal affinity. *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing - CSCW 16*.

Logg, J. M., Minson, J. A., and Moore, D. A. (2019). Algorithm appreciation: People prefer algorithmic to human judgment. *Organizational Behavior and Human Decision Processes*, 151:90–103.

Luo, X., Qin, M. S., Fang, Z., and Qu, Z. (2021). Artificial Intelligence Coaches for Sales Agents: Caveats and Solutions. *Journal of Marketing*, 85(2):14–32.

Luo, X., Tong, S., Fang, Z., and Qu, Z. (2019). Frontiers: Machines vs. humans: The impact of artificial intelligence chatbot disclosure on customer purchases. *Marketing Science*.

Malle, B. F. and Pearce, G. E. (2001). Attention to behavioral events during interaction:

Two actor-observer gaps and three attempts to close them. *Journal of Personality and Social Psychology*, 81(2):278–294.

Malone, P. B. (1980). Humor: A double-edged tool for todays managers? *The Academy of Management Review*, 5(3):357.

Malthouse, E. C., Haenlein, M., Skiera, B., Wege, E., and Zhang, M. (2013). Managing customer relationships in the social media era: Introducing the social crm house. *Journal of Interactive Marketing*, 27(4):270–280.

Mann, H. B. and Whitney, D. R. (1947). On a test of whether one of two random variables is stochastically larger than the other. *The Annals of Mathematical Statistics*, 18(1):50–60.

March, C. (2019). The Behavioral Economics of Artificial Intelligence: Lessons from Experiments with Computer Players. *CESifo Working Paper*, (7926).

Mathur, M. B. and Reichling, D. B. (2016). Navigating a social world with robot partners: A quantitative cartography of the uncanny valley. *Cognition*, 146:22–32.

McCabe, K., Houser, D., Ryan, L., Smith, V., and Trouard, T. (2001). A functional imaging study of cooperation in two-person reciprocal exchange. *Proceedings of the National Academy of Sciences of the United States of America*, 98(20):11832–11835.

Meuter, M. L., Ostrom, A. L., Roundtree, R. I., and Bitner, M. J. (2000). Self-service technologies: Understanding customer satisfaction with technology-based service encounters. *Journal of Marketing*, 64(3):50–64.

Miller, A. R. and Tucker, C. (2012). Active social media management: The case of health care. *SSRN Electronic Journal*.

Miller, C. H., Lane, L. T., Deatrick, L. M., Young, A. M., and Potts, K. A. (2007). Psychological reactance and promotional health messages: The effects of controlling language, lexical concreteness, and the restoration of freedom. *Human Communication Research*, 33(2):219–240.

Mone, G. (2016). The edge of the uncanny. *Communications of the ACM*, 59(9):17–19.

Moon, Y. (1999). The effects of physical distance and response latency on persuasion in computer-mediated communication and human–computer communication. *Journal of Experimental Psychology: Applied*, 5(4):379–392.

Moon, Y. (2000). Intimate exchanges: Using computers to elicit self-disclosure from consumers. *Journal of Consumer Research*, 26(4):323–339.

Moon, Y. and Nass, C. (1996). How "Real" Are Computer Personalities? Psychological Responses to Personality Types in Human-Computer Interaction. *Communication Research*, 23(6):651–674.

Moore, S., Zemack-Rugar, Y., and Fitzsimons, G. J. (2014). Buy now! how brand relationships influence consumer responses to imperative advertising. *ACR North American Advances*.

Moretti, L. and Pellegrino, G. D. (2010). Disgust selectively modulates reciprocal fairness in economic interactions. *Emotion*, 10(2):169–180.

Morkes, J., Kernal, H. K., and Nass, C. (1999). Effects of humor in task-oriented human-computer interaction and computer-mediated communication: A direct test of srct theory. *Human–Computer Interaction*, 14(4):395–435.

Möttönen, T., Hannonen, P., Leirisalo-Repo, M., Nissilä, M., Kautiainen, H., Korpela, M., Laasonen, L., Julkunen, H., Luukkainen, R., Vuori, K., et al. (1999). Comparison of combination therapy with single-drug therapy in early rheumatoid arthritis: a randomised trial. *The Lancet*, 353(9164):1568–1573.

Nass, C. and Lee, K. M. (2001). Does computer-synthesized speech manifest personality? experimental tests of recognition, similarity-attraction, and consistency-attraction. *Journal of Experimental Psychology: Applied*, 7(3):171–181.

Nass, C. and Moon, Y. (2000). Machines and mindlessness: Social responses to computers. *Journal of Social Issues*, 56(1):81–103.

Nass, C., Steuer, J., and Tauber, E. R. (1994). Computers are social actors. *Conference companion on Human factors in computing systems - CHI 94*.

Niculescu, A., Dijk, B. V., Nijholt, A., Li, H., and See, S. L. (2013). Making social robots more attractive: The effects of voice pitch, humor and empathy. *International Journal of Social Robotics*, 5(2):171–191.

Nowak, K. L. and Biocca, F. (2003). The effect of the agency and anthropomorphism on users sense of telepresence, copresence, and social presence in virtual environments. *Presence: Teleoperators and Virtual Environments*, 12(5):481–494.

Oliver, R. L. (1977). Effect of expectation and disconfirmation on postexposure product evaluations: An alternative interpretation. *Journal of Applied Psychology*, 62(4):480–486.

Pan, S. J. and Yang, Q. (2010). A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1345–1359.

Powers, A. and Kiesler, S. (2006). The advisor robot. *Proceeding of the 1st ACM SIGCHI/SIGART conference on Human-robot interaction - HRI 06*.

Promberger, M. and Baron, J. (2006). Do patients trust computers? *Journal of Behavioral Decision Making*, 19(5):455–468.

Sah, Y. J. and Peng, W. (2015). Effects of visual and linguistic anthropomorphic cues on social perception, self-awareness, and information disclosure in a health website. *Computers in Human Behavior*, 45:392–401.

Sanfey, A. G., Rilling, J. K., Aronson, J. A., Nystrom, L. E., and Cohen, J. D. (2003). The neural basis of economic decision-making in the ultimatum game. *Science*, 300(5626):1755–1758.

Schanke, S. Wagle, M. R. G. A. G. (2017). The style and content of posts - a study of corporate engagement on facebook. *working paper*.

Schweitzer, M. E., Hershey, J. C., and Bradlow, E. T. (2006). Promises and lies:

Restoring violated trust. *Organizational Behavior and Human Decision Processes*, 101(1):1–19.

Searle, J. R. (1975). A taxonomy of illocutionary acts.

Singh, R., Tay, Y. Y., and Sankaran, K. (2017). Causal role of trust in interpersonal attraction from attitude similarity. *Journal of Social and Personal Relationships*, 34(5):717–731.

Sproull, L., Subramani, M., Kiesler, S., Walker, J., and Waters, K. (1996). When the interface is a face. *Human-Computer Interaction*, 11(2):97–124.

Sun, T., Yuan, Z., Li, C., Zhang, K., and Xu, J. (2020). The value of personal data in internet commerce: A high-stake field experiment on data regulation policy. *SSRN Electronic Journal*.

Swearingen, K. and Sinha, R. (2001). Beyond Algorithms : An HCI Perspective on Recommender Systems. *ACM SIGIR 2001 Workshop on Recommender Systems (2001)*, pages 1–11.

Tambe, P., Cappelli, P., and Yakubovich, V. (2019). Artificial intelligence in human resources management: Challenges and a path forward. *California Management Review*, 61(4):15–42.

Taylor, S. (1994). Waiting for service: The relationship between delays and evaluations of service. *Journal of Marketing*, 58(2):56.

Teubner, T., Adam, M., and Riordan, R. (2015). The impact of computerized agents on immediate emotions, overall arousal and bidding behavior in electronic auctions. *Journal of the Association for Information Systems*, 16(10):838–879.

Thaler, R. H. (2018). Behavioral economics: Past, present, and future. *Revista de Economia Institucional*, 20(38):9–43.

Torta, E., Dijk, E. V., Ruijten, P. A. M., and Cuijpers, R. H. (2013). The ultimatum game as measurement tool for anthropomorphism in human–robot interaction. *Social Robotics Lecture Notes in Computer Science*, page 209–217.

Trainer, T., Taylor, J. R., and Stanton, C. J. (2020). *Choosing the Best Robot for the Job: Affinity Bias in Human-Robot Interaction*, volume 12483 LNAI. Springer International Publishing.

Verhagen, T., Nes, J. V., Feldberg, F., and Dolen, W. V. (2014). Virtual customer service agents: Using social presence and personalization to shape online service encounters. *Journal of Computer-Mediated Communication*, 19(3):529–545.

Walther, J. B. (1992). Interpersonal effects in computer-mediated interaction. *Communication Research*, 19(1):52–90.

Walther, J. B. and Tidwell, L. C. (1995). Nonverbal cues in computer-mediated communication, and the effect of chronemics on relational communication. *Journal of Organizational Computing*, 5(4):355–378.

Wang, F.-Y., Carley, K. M., Zeng, D., and Mao, W. (2007a). Social computing: From social informatics to social intelligence. *IEEE Intelligent systems*, 22(2):79–83.

Wang, L. C., Baker, J., Wagner, J. A., and Wakefield, K. (2007b). Can a retail web site be social? *Journal of Marketing*, 71(3):143–157.

Weiss, A. and Bartneck, C. (2015). Meta analysis of the usage of the godspeed questionnaire series. *2015 24th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*.

Wilson, H., Daugherty, P., and Bianzino, N. (2017). When ai becomes the new face of your brand. *Harvard Business Review*, 27.

Wirtz, J., Patterson, P. G., Kunz, W. H., Gruber, T., Lu, V. N., Paluch, S., and Martins, A. (2018). Brave new world: service robots in the frontline. *Journal of Service Management*, 29(5):907–931.

Xu, K. and Lombard, M. (2017). Persuasive computing: Feeling peer pressure from multiple computer agents. *Computers in Human Behavior*, 74:152–162.

Ye, S., Gao, G. G., and Viswanathan, S. (2014). Strategic behavior in online reputation systems: Evidence from revoking on ebay. *MIS Quarterly*, 38(4):1033–1056.

Yeomans, M., Shah, A., Mullainathan, S., and Kleinberg, J. (2019). Making sense of recommendations. *Journal of Behavioral Decision Making*, 32(4):403–414.

Yeong Tan, D. T. and Singh, R. (1995). Attitudes and Attraction: A Developmental Study of the Similarity-Attraction and Dissimilarity-Repulsion Hypotheses. *Personality and Social Psychology Bulletin*, 21(9):975–986.

Yoon, E., Henry, M., Goodman, N., and Frank, M. (2016). Talking with tact: Polite language as a balance between kindness and informativity. *Proceedings of the 38th Annual Conference of the Cognitive Science Society*.

Zemack-Rugar, Y., Moore, S. G., and Fitzsimons, G. J. (2017). Just do it! why committed consumers react negatively to assertive ads. *Journal of Consumer Psychology*, 27(3):287–301.

Appendix A

Appendix

A.1 Multicollinearity

We also assessed whether our regression results are subject to multicollinearity issues. As one might expect, the interaction between the *Treatment Count* and *Cash Offer* exhibits a relatively high correlation with the constituent terms, and produces a relatively high variance inflation factor (VIF) in turn. However, it should be kept in mind that high VIFs are not typically problematic when they result from correlations between interaction terms and their constituent variables. To demonstrate this in our setting, we apply the residual-centering approach of (Lance 1988). The resulting regression yields very similar results to the baseline model, and the VIF values (reported along side the centered model values) are well within normal thresholds (see Table A.1). Ultimately, we conclude that collinearity is not influencing the results.

Table A.1: Treatment Count Model (LPM; DV = Convert)

Variable	Original Model	Residual Centering	VIF	1/VIF
1 $Treatment$	0.076** (0.032)	.0756** (.0317)	2.46	0.407
2 $Treatments$	0.060** (0.030)	.0603** (.0297)	2.46	0.4062
3 $Treatments$	0.109** (0.049)	.1213** (.0508)	1.85	0.5408
1 $Treatment \cdot Cash\,Offer$	0.052 (0.064)	.0115 (.0162)	3.42	0.2926
2 $Treatments \cdot Cash\,Offer$	0.086 (0.069)	.0216 (.0175)	3.43	0.2916
3 $Treatments \cdot Cash\,Offer$	0.211** (0.087)	.0509** (.0121)	2.46	0.4072
$Cash\,Offer$	−0.058 (0.057)	.0062 (.0057)	1.01	0.9936
Intercept	0.017 (0.017)	.0174 (.0169)	Mean VIF	2.44
$Observations$	323	323		
R^2	0.037	0.036		
F	3.85*** $(7, 316)$	3.82*** $(7, 315)$		

Note: Robust SEs; ** $p < 0.05$, * $p < 0.10$.

A.2 Estimator Choice

One possible concern with our results is that they are somehow dependent upon bias or inconsistency of the LPM (Horrace and Oaxaca 2006). It is important to note, however, first, that the typical concerns with bias and inconsistency of OLS and binary outcomes are not applicable to experimental treatment impact evaluations (Deke et al. 2014). Second, even in observational data, Horrace and Oaxaca (2006) have shown that the bias underlying LPMs is unlikely to be severe when the vast majority of predicted values a resulting model yields fall entirely within the 0-1 range. In the event that any predicted values do lie outside the feasible range, those authors propose the application of a trimming estimator. This estimator is a standard LPM that simply omits those observations holding infeasible predicted values. Employing this procedure notably only results in our excluding 5 observations from the original sample and, as can be seen in Table A.2, our coefficients remain essentially unchanged.

Although a Logistic regression is often viewed as preferable when dealing with binary outcomes, because it has the desirable property of constraining predicted values to lie within the 0-1 interval, it is important to keep in mind that this estimator also has the *undesirable* property of yielding coefficients that are difficult to understand or interpret. This is true for two reasons. First, logistic regression

Table A.2: Trimmed OLS (LPM; DV = Convert)

Coefficient	Model (1)
1 $Treatment$	0.0870^{***} (0.0279)
2 $Treatments$	0.0711^{***} (0.0259)
3 $Treatments$	0.1171^{**} (0.0461)
1 $Treatment \cdot Cash\ Offer$	0.01938 (0.0217)
2 $Treatments \cdot Cash\ Offer$	0.0295 (0.0240)
3 $Treatments \cdot Cash\ Offer$	0.0295^{**} (0.0230)
$Cash\ Offer$	-0.0223 (0.0217)
Intercept	0.0054 (0.0075)
$Observations$	318
R^2	0.035
F	3.82^{***} $(7, 310)$

Note: Robust SEs; ** $p < 0.05$, * $p < 0.10$.

deals with odds ratios, which often lack straightforward intuition, given their multiplicative nature. Second, the coefficients and standard errors associated with interaction terms in Logistic Regression are not directly interpretable (Ai and Norton 2003). That said, we also estimated a logistic regression model, the results of which are presented below in Table A.3. As can be seen, these results are qualitatively similar to those reported elsewhere, in terms of sign and statistical significance of the estimated coefficients.

Table A.3: Logit (DV = Convert)

Coefficient	Model (1)
1 $Treatment$	3.879^{***} (1.083)
2 $Treatments$	3.641^{***} (1.099)
3 $Treatments$	3.813^{***} (1.168)
1 $Treatment \cdot Cash\,Offer$	1.127^{***} (0.337)
2 $Treatments \cdot Cash\,Offer$	1.268^{***} (0.361)
3 $Treatments \cdot Cash\,Offer$	1.533^{***} (0.351)
$Cash\,Offer$	-1.161^{***} (0.324)
Intercept	-6.168 (1.036)
$Observations$	323
$Wald\ Chi^2$	29.14^{***}
$Pseudo\ R^2$	0.0659

Note: Robust SEs; ** $p < 0.05$, * $p < 0.10$.

A.3 Randomization Checks

In this section, we report additional randomization checks, evaluating the orthogonality of cash offer and treatment assignment to one another, as well as between cash offer assignment and the clothing items that a subject brought to the buy-back procedure. Table A.4 shows the pairwise comparisons of the average cash offer assigned between alternative conditions, defined in terms of the number of treatments assigned. In Table A.5, we also report pairwise comparisons between each of the eight individual conditions, defined in terms of the unique combination of treatments assigned. All mean differences are statistically insignificant at the p<.05 level. These null results indicate that cash offer assignment was orthogonal to anthropomorphism treatment assignment.

Table A.4: Pairwise Comparisons cash offer and Number of Treatments

Test	**Condition Comparison**	***t*-stat**	***p*-value**
1	1 Treatment vs 2 Treatments	-0.228	0.820
2	1 Treatment vs 3 Treatments	-0.700	0.485
3	2 Treatments vs 3 Treatments	-0.527	0.600
4	1 Treatment vs 0 Treatments	-0.957	0.340
5	2 Treatments vs 0 Treatments	-1.117	0.266
6	3 Treatments vs 0 Treatments	-1.405	0.164

Finally, evaluating the correlation between the cash offer a subject was assigned and the number of clothes he or she was was seeking to sell (conditional on their progressing beyond the cash offer offer stage of the conversation), we again observed a statistically insignificant relationship ($p > 0.05$).

Table A.5: Pairwise Comparisons Cash Offer and Combination of Treatments

Test	**Condition Comparison**	***t*-stat**	***p*-value**
1	SP vs D	.093	0.93
2	SP vs H	0.172	0.86
3	SP vs SP & D	0.107	0.92
4	SP vs D & H	0.02	0.98
5	SP vs SP & H	-0.759	0.45
6	SP vs SP & D & H	0.554	0.58
7	SP vs Control	0.808	0.42
6	D vs H	0.052	0.96
7	D vs SP & D	0.00	1.00
8	D vs D & H	0.062	0.95
9	D vs SP & H	.771	0.44
10	D vs SP & D & H	-0.582	0.56
11	D vs Control	.093	0.54
12	H vs SP & D	-0.059	0.95
13	H vs D & H	-0.1370	0.89
14	H vs SP & H	-1.070	0.29
15	H vs SP & D & H	-0.777	0.44
16	H vs Control	0.445	0.77
17	SP & D vs D & H	-0.071	0.94
18	SP & D vs SP & H	-0.883	0.38
19	SP & D vs SP & D & H	-0.668	0.51
20	SP & D vs Control	0.710	0.48
21	D & H vs SP & H	-0.707	0.48
22	D & H vs SP & D & H	-0.505	0.61
23	D & H vs Control	0.717	0.48
24	SP & H vs SP & D & H	0.178	0.86
25	SP & H vs Control	1.717	0.09
26	SP & D & H vs Control	1.398	0.17

SP= Social Presence, D=Delay, H=Humor

A.4 Individual Treatments

The focus of the study is on the affects of anthropomorphism. However, from a practical standpoint, it is likely useful to also understand which of our treatments are most effective, individually. We therefore conducted additional analyses and another experiment on Amazon Mechanical Turk, aimed at addressing this question.

First, we report on our Turk experiment. In this experiment, we evaluated the desirability of individual treatment interventions in terms of subjects' reported perception of chatbot likeability. We limit this analysis to an Appendix, because it is not altogether clear whether responses from this artificial setting, in which subjects are paid to participate, would necessarily mirror results obtained in a field setting, wherein individuals organically opt into chatbot interactions. That said, results of this analysis may provide a useful indication of which anthropomorphic interventions may be particularly useful in practice.

We recruited 426 subjects on Mechanical Turk to interact with four versions of our chatbot, assigned at random: i. control, ii. social presence, iii. delay and iv. humor. We limited participation such that a given Turker could complete the HIT exactly once, to avoid concerns about interference and cross-over across conditions. After Turkers interacted with the their assigned chatbot, they were asked to rate the chatbot on the Mathur and Reichling (2016) enjoyable/unpleasant scale, which ranges from -100 to 100. We find that the humor treatment yields a significantly larger, positive effect than either the control or the delay treatment - Mann-Whitney U tests indicate statistical significance at conventional levels (p ¡ 0.05). We provide a visual depiction of the average likeability report by experimental condition in Figure A.1.

We also note here that the delay treatment yields significantly lower likeability than the control condition in this setting. As noted earlier, it is not clear whether this finding would also apply to our field setting. It should be kept in mind that crowd-workers are paid for their time. As such, our delay treatment in this setting not only manipulates anthropomorphism; it also implies that turkers are earning a lower effective wage. Moreover, we would note that we also do not account for interactions between different anthropomorphic treatments here. As such, it remains possible that delay can have a strictly positive, amplifying effect, as long as it is implemented in tandem with other anthropomorphic treatments.

Next, we revisited the data from our initial field experiment. A natural approach to consider is to simply remove our *Treatment Count* dummies and replace them with individual treatment dummies, along with all possible interactions. Unfortunately, such a model is under-powered, and yields a statistically insignificant overall model fit. Accordingly, we considered a simpler regression specification,

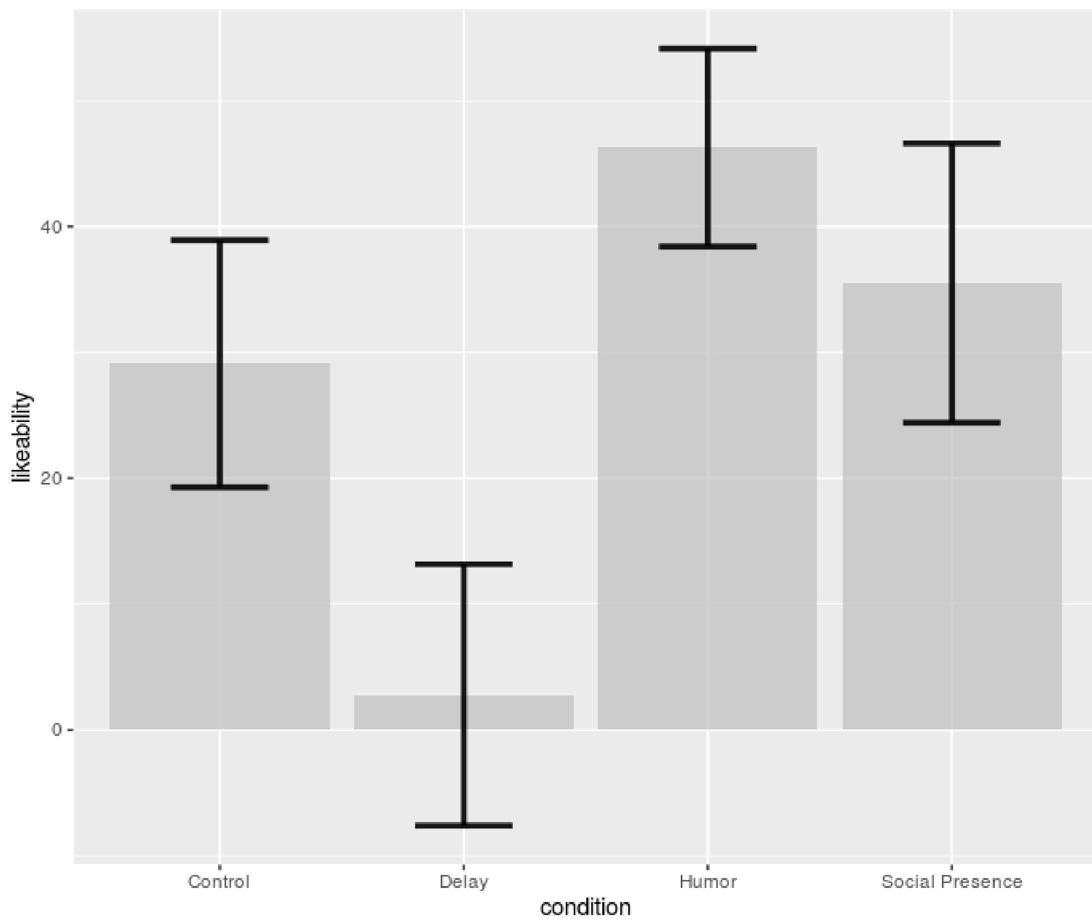

Figure A.1: Perceived Likeability - Individual Anthropomorphic Treatments which ignores interactions between anthropomorphism treatments, and merely seeks to assess average main effects of each individual treatment, as well as their offer interactions. The model remains valid, of course, because all treatments and offer manipulations were varied exogenously.

Notably, this new model, estimated using the Horrace and Oaxaca (2006) trimming estimator, is statistically significant overall. The models yields an F-stat of 2.70 (7, 301), implying a p-value of 0.01 for overall model fit. The model results appear below in Table A.6. We observe results consistent with those seen in our Amazon Mechanical Turk study, above. That is, *Humor* has a significant, positive effect on conversion, whereas the coefficients on our two other interventions are null. Further, the effect of *Humor* is significantly larger than the effect of *Delay* (p = 0.07). Additionally, we see that *Delay* has a statistically significant interaction with *Cash Offer*, suggesting it has a particular influence on offer sensitivity. Of course, these results are far from conclusive; additional work should be pursued to identify the ideal anthropomorphic interventions for retail settings.

Table A.6: LPM (DV = Convert)

Coefficient	Trimmed LPM (1)
$Delay$	-0.012 (0.030)
$Humor$	0.067^{**} (0.032)
$SocialPresence$	0.018 (0.031)
$Delay \cdot Cash\ Offer$	0.156^{**} (0.054)
$Humor \cdot Cash\ Offer$	-0.002 (0.051)
$SocialPresence \cdot Cash\ Offer$	0.091 (0.053)
$Cash\ Offer$	-0.103 (0.057)
$Observations$	309
$F - stat$	$2.70^{*}(7, 301)$
R^2	0.052

Note: Robust SEs;** $p < 0.05$, * $p < 0.10$.